Heine and Robitzsch
Large-scale Assessments in Education (2022) 10:10
https://doi.org/10.1186/s40536-022-00129-5

Large-scale Assessments in Education

RESEARCH Open Access

Evaluating the effects of analytical decisions in large-scale assessments: analyzing PISA mathematics 2003-2012

Jörg-Henrik Heine[1*] and Alexander Robitzsch[2]

*Correspondence:
joerg.heine@tum.de

[1] Center for International Student Assessment (ZIB), Department of Educational Sciences, Technical University of Munich (TUM), School of Social Sciences and Technology, Arcisstrasse 21, 80333 Munich, Germany
[2] IPN Leibniz Institute for Science and Mathematics Education, Center for International Student Assessment (ZIB), Olshausenstraße 62, 24118 Kiel, Germany

Abstract

Research question: This paper examines the overarching question of to what extent different analytic choices may influence the inference about country-specific cross-sectional and trend estimates in international large-scale assessments. We take data from the assessment of PISA mathematics proficiency from the four rounds from 2003 to 2012 as a case study.

Methods: In particular, four key methodological factors are considered as analytical choices in the rescaling and analysis of the data: (1) The selection of country sub-samples for item calibration differing at three factor levels. (2) The item sample refering to two sets of mathematics items used within PISA. (3) The estimation method used for item calibration: marginal maximum likelihood estimation method as implemented in R package TAM or an pairwise row averaging approach as implemented in the R package pairwise. (4) The type of linking method: concurrent calibration or separate calibration with successive chain linking.

Findings: It turned out that analytical decisions for scaling did affect the PISA outcomes. The factors of choosing different calibration samples, estimation method and linking method tend to show only small effects on the country-specific cross-sectional and trend estimates. However, the selection of different link items seems to have a decisive influence on country ranking and development trends between and within countries.

Keywords: Large-scale assessment, PISA, Mathematics item sampling, Trend estimate, Estimation method, linking, Scaling, Cross-sectional estimate

Introduction

International large-scale assessments (ILSA) as comparative education studies have gained prominence in recent decades in global, national, and even local education debates (UNESCO, 2019). Such recent studies as Trends in International Mathematics and Science Study (TIMSS, e.g., Martin et al., 2020), Progress in International Reading Literacy Study (PIRLS, e.g., Martin et al., 2017), and the Programme for International Student Assessment (PISA) date back in their origins to the 1950s and 1960s in which education became an active field of inquiry for all the social sciences and

thus comparative education began to make increasing use of the more mature and developed social science methods (Anderson, 1961; Henry, 1973). Since these early origins and the associated discussions in comparative educational science about different methodological approaches to the field (see. Henry 1973), methodology typical for each comparative educational study has evolved in the last few decades. For two decades, the OECD's studies for the Programme for International Student Assessment (PISA, e.g., OECD et al., 2009, OECD, 2012, 2014, 2017, 2021a) and the reporting of their results in a recurring three-year cycle have been a notable event in media coverage of widely discussed issues in secondary education (e.g., Grek, 2009). In addition to this media presence of international educational tests such as PISA, the comparison of one's own country's performance with that of countries with higher scores in particular not infrequently tempts political decision-makers to draw educational policy conclusions on this supposedly rock-solid empirical basis and to take ad hoc remedial actions that not infrequently turn out to be misleading in the long run (e.g., Singer & Braun, 2018).

Although from a methodological perspective, the study design of PISA, as well as TIMSS and PIRLS, cannot be termed panel or longitudinal studies, questions of development trends are becoming increasingly prominent in the reporting and reception of ILSA results. They might be used to legitimize national educational reforms (e.g., Fischman, 2019; Johansson, 2016; Grek, 2009). For PISA such trend observations are vindicated on the one hand despite cross-sectional but representative sampling of fifteen-year-old students at the respective survey period and on the other hand on the relative continuity about the type of data collected. The recurrent PISA results in the three core domains of reading, mathematics and science are therefore regarded as comparative trend indicators of the performance of the educational systems in the respective participating countries. Apart from the aspect that the resulting competitive horse race communication can be criticized as such (see Ertl, 2020), the question of how to methodically underpin the trend statements is therefore becoming increasingly important (see Singer & Braun, 2018).

The present article aims to investigate to what extent different analytical decisions regarding item calibration, proficiency scaling and linking of the single ILSA rounds may lead to different statements concerning development trends within and between the participating countries. Specifically, using PISA data collected in the past 2003 to 2012 rounds, we examine how different analytic choices in international comparative assessment might contribute to contrasting conclusions about the country's mean differences in mathematics literacy when examined cross-sectionally and by trend.

In detail, these analytical choices relate to the type of selection of country sub-samples for item calibration, considering three different options as factor levels. Second, the selection of the (link) item sample refers to two different sets of items used within PISA from 2003 up to 2012. Third, the estimation method of item calibration is varied by applying two different types of estimation methods. Furthermore, we consider two types of linking methods as a basis for the cross-sectional country comparisons and trend analyses. We consider these different analytical choices as potential sources to increase the methodological variance in scaling and data analysis, leading to statements deviating from the official reporting concerning the cross-sectional and trend estimates in PISA.

For this purpose, we organized the present article as follows. First, an overview of the official methods for scaling the cognitive data in the PISA large-scale assessment (LSA) is given. The focus here is on the model and estimation method used for item calibration and scaling as well as the principle for linking different PISA rounds as it has been applied in PISA from 2003 up to 2012. We supplement this with some selected examples of empirical findings and theoretical considerations from the literature that critically address this 'official' methodology used in reporting PISA outcomes so far. In turn, we investigate a strategy for reanalyzing the PISA database covering the cross-sectional assessment data from the beginning in the year 2003 up to the last paper-based PISA assessment in 2012. In the methods section of this paper, we also describe the extensive data preparation procedures in the form of a brief summary. This process of adapting and harmonizing the single cross-sectional data sets, which precedes the actual analysis, is necessary because the coding of student responses, the different naming of specific items with the same content, and the general handling of the data have been subject to numerous changes over the four PISA rounds. However, this adaptation and harmonization is an essential prerequisite for the reanalysis of trend and cross-sectional analyses and may pose a potential burden to other researchers that should not be underestimated when dealing with historical data.

In addition, the four analytical decisions considered in the analyses are presented. Based on the findings from the literature, these analytic decisions refer to the selection of the (link) items, the selection of the calibration sample(s), the estimation method utilized for item calibration, as well as the way of linking different PISA rounds. Finally, the results are discussed against the backdrop of the increasing influence of PISA results on policy decisions and longitudinal trend statements on the development of educational systems.

Principles in OECD calibration, scaling, linking and trend reporting for PISA 2003 – 2012

Since its first implementation in 2000, the analysis of the data collected in PISA has been based on scaling models from the item response theory (IRT). For the PISA rounds 2000 up to 2012, the IRT base model used for item calibration and scaling principle is the partial credit model (PCM; Masters, 1982), which is an extension of the Rasch model (Rasch, 1960) for polytomous item responses. The probability of an answer to item i with K_i categories in category k ($k = 0, \ldots, K_i$) is given by

$$\text{Prob}(X_i = k|\theta) = \frac{\exp(k\theta + d_{ik})}{\sum_{h=0}^{K_i} \exp(h\theta + d_{ih})}\,, \tag{1}$$

where θ is the unidimensional ability variable and d_{ik} is the difficulty of the k'th 'step' of item i (see Masters, 1982 p. 172), with $d_{i0} = 0$, standardized at the sum over all categories of the exponent of the difference of θ and d_{ih} with $h = 0, \ldots, K_i$. The specific model used for the multidimensional IRT scaling of the PISA domains was the mixed coefficients multinomial logit model (MCMLM; Adams et al., 1997), which can be seen as a generalization of the unidimensional PCM to model student ability in D correlated dimensions $\theta_1, \ldots, \theta_D$. In the MCMLM (see Adams et al., 1997), the item response of item i in category k is modeled as

$$\text{Prob}(X_i = k, \boldsymbol{A}, \boldsymbol{B}|\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{b}_{ik}\boldsymbol{\theta} + \boldsymbol{a}_{ik}\boldsymbol{\xi})}{\sum_{h=0}^{K_i} \exp(\boldsymbol{b}_{ih}\boldsymbol{\theta} + \boldsymbol{a}_{ih}\boldsymbol{\xi})}, \tag{2}$$

where $\boldsymbol{\xi}$ is the vector of estimated item parameters, which after reparametrization are the basis for the (mean) item difficulty δ_i (see Eq. 3 below), and known design matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ containing all vectors $\boldsymbol{a}_{ik}$ and $\boldsymbol{b}_{ik}$ ($i = 1, \ldots, I$, $k = 0, \ldots, K_i$), respectively. For the complete definition of the population model in PISA, the distribution of the vector of latent variables $\boldsymbol{\theta}$ is modeled by a multivariate normal density $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.

In PISA, this model was used for the official reporting of all rounds from 2000 to 2012 in two steps, preceded by a national calibration carried out separately for each country. The preceding national calibration step served to monitor the quality of the data and to provide a basis for deciding how to treat each item in each country. In some cases, this could lead to the removal of an item from the PISA reporting if it had poor psychometric properties in more than ten countries (a "dodgy" item, OECD, 2014 p. 148). First, in the international item calibration step, often referred to as international scaling in OECD technical reports, the item parameters are determined across countries, with the underlying response data consisting of 500 randomly selected students from each OECD country sample serving as an international calibration sample. In the second step, the student abilities were estimated by including an additional conditioning component in the scaling model. For this, $\boldsymbol{\mu}$ from the population model is replaced by a regression component $\boldsymbol{y}_n\boldsymbol{\beta}$ where $\boldsymbol{y}_n$ is a vector for student n containing additional student information from the background questionnaire variables, and $\boldsymbol{\beta}$ is the corresponding matrix of regression coefficients. Note that in this latent regression model, the student abilities are not estimated directly, but a posterior distribution for the latent variable is specified from which plausible values (PV) are drawn. This principle of latent regression IRT modeling using auxiliary (student) information to estimate population characteristics is described by Mislevy et al., (1992) and is based on the principles of handling missing data by multiple imputation (Rubin 1987), adapted for proficiency estimation based on data resulting from booklet designed proficiency tests (see also Mislevy, 1991).

Based on such a cross-sectional calibration and scaling approach, the successive chain linking for trend analysis of PISA results across different rounds requires the existence of common items from earlier assessment cycles (see, e.g., Mazzeo and Von Davier, 2013). Typically, the following six steps were performed for linking proficiency measures between different PISA rounds until 2012 (OECD, 2014). In a first step, a calibration of the item difficulties was performed using the calibration sample from the current PISA round, as already mentioned above. In the second step, the obtained item difficulties are transformed with a constant such that the mean values of the item parameters of the common items are set equal to those from the previous round or the round to be linked. In the third step, the data set for all OECD countries in PISA 2012 is scaled twice – once with all items of the respective competence domain and once with the link items only. In the fourth step, for the sample of OECD countries, the difference between the two scalings is removed by applying an additional linear transformation that accounts for differences in means and variances in

the two scalings (Gebhardt and Adams, 2007). This is followed in step five by estimating the person parameters (ability) for the current PISA round, which are anchored to the initial item parameters (first calibration step). Finally, the person parameters are transformed using the calculated transformation constants from steps two and four in the last step. As a result of such a linking approach (e.g., Dorans et al.,2007), proficiency estimates from different rounds can be directly compared on the same metric (Mazzeo and Von Davier, 2013).

The official PISA methodology address the uncertainty, which is associated with the round-wise calibration, when comparing different PISA rounds by taking into account so-called link errors. The basic idea behind the calculation of the link errors in PISA (up to the 2012 round) is to consider the differential functioning of the common items (DIF) across the PISA rounds to be compared (OECD, 2014), as it results from the respective international item calibrations from each single PISA round. Thus, in order to calculate the link error for the PISA round 2006 compared to the previous round 2003, first, the differences $\widehat{\delta}_{i,2006} - \widehat{\delta}_{i,2003}$ of the respective IRT estimated item difficulties $\widehat{\delta}_i$ of a set of I_0 common link items can be computed. Under the assumption that the used link items represent a random sample of all possible link items, the link error $LE_{2003,2006}$ for trend estimates for country means was then estimated as follows:

$$LE_{2003,2006} = \sqrt{\frac{1}{I_0}\sum_{i=1}^{I_0}\left(\widehat{\delta}_{i,2006} - \widehat{\delta}_{i,2003}\right)^2}\,. \tag{3}$$

This basic principle of linking presented here was retained for all further rounds up to 2012, whereby, however, the clustering of the items in individual units (item stems) and, as an additional item weighting factor, the fact that items with polytomous response formats have a greater influence on the competence scale score than dichotomous ones were additionally taken into account (see OECD, 2014 pp.160). The standard error $SE_{2003,2006}$ for a difference of the two country means from the PISA rounds 2003 and 2006 is determined by the two round-specific components σ_μ and the link component:

$$SE_{2003,2006} = \sqrt{SE_{2003}^2 + SE_{2006}^2 + LE_{2003,2006}^2}\,, \tag{4}$$

where SE_{2003} and SE_{2006} denote standard errors for the country means in PISA 2003 and 2006, respectively. Further detailed information and a formal description of the official procedure for determining link errors in the PISA rounds up to 2012 can be found in the technical report on PISA 2012 (see OECD, 2014 pp.159–163) as well as in the Annex A5 of the PISA results volume I (OECD, 2014).

The link errors determined in this way are then, for example, taken to supplement the standard errors of the country means to be compared in analyses of mean differences between countries. It can therefore be said that at its core, PISA uses a special case of the variance component model (see Robitzsch & Lüdtke, 2019) to determine composite standard errors. In the official analysis and reporting of PISA, such a model took into account only the variance component of the international item parameters across single PISA rounds to be compared and in addition to this DIF, takes into account the clustering and the response format of the items (OECD, 2014).

These analytical decisions and official procedures for calibrating, scaling, linking and reporting PISA results, briefly outlined here, have inspired some critical theoretical discussions and methodological research, which in turn evoked criticism about the PISA methodology. In the following section, we briefly review some key aspects of this criticism.

On analytical decisions in large-scale assessments

The analytical principles outlined in the previous section and the resulting official methodological procedures for calibrating, scaling, linking, and evaluating PISA results have attracted various criticisms over time. These refer to different aspects of the applied methodology, each supported by recent empirical findings or simulation outcomes (e.g., Rutkowski et al., 2016; Rutkowski, 2014; Rutkowski, 2011; von Davier et al., 2019, Robitzsch & Lüdtke, 2019; Rutkowski et al., 2019). For example, studies such as (Rutkowski, 2014) suggest that using a background model for latent regression, besides its theoretically derived advantages (see Mislevy et al., 1992), can also be seen as an additional source of error variance to an uncertain extent. Specifically, (Rutkowski, 2014) shows that the misclassification of subjects based on deficient background information results in mean differences of groups being significantly underestimated or overestimated, which can also be interpreted with an under- or overestimation of variance in relation to the entire population. Thus, although using a background model in the scaling of ability estimates is currently a standard evaluation procedure in many large-scale assessments, this approach can also be criticized. This criticism is usually based on the suspected and sometimes empirically proven poor quality of the questionnaire data used in such latent regression models (see, e.g., Hopfenbeck & Maul, 2011). Typically, the criticized poor quality of the questionnaire data results from the high proportion of missing values (e.g., Rutkowski, 2011; Grund et al., 2021). In contrast, almost paradoxically, the introduction of the latent regression model is motivated precisely by the targeted increase in the estimation accuracy of the model parameters of the response model against the background of missing values by rotated booklet designs, as well as missing student responses in the cognitive assessment materials (see, e.g., Mislevy et al., 1992; Rubin, 1987; Mislevy, 1991). However, in the current practice of scaling PISA data using latent regression models, the necessary prerequisite of complete background questionnaire data is realized by the quite weak missing indicator method (MIM; Cohen & Cohen, 2003), which has been shown to be inadequate and prone to bias if missingness in the background variables is not missing completely at random (e.g., Schafer & Graham, 2002; Grund et al., 2021). The method of parameter estimation typically associated with the latent regression models is marginal maximum likelihood (MML) estimation. The efficiency of MML estimation based on this full information approach is founded on the theoretical assumption that with an asymptotic infinite size, no other estimator provides parameter estimates with smaller variances (e.g., Forero & Maydeu-Olivares, 2009). Under the assumptions of multivariate normality (but see Xu & von Davier, 2008 for modeling deviations from normality) and a correctly specified model, the latent variable model parameters are consistently estimated by simultaneous equation methods, for instance, full information maximum likelihood (FIML) (Lance et al., 1988). However, (Lance et al., 1988) pointed out that estimation methods with complete information may also have

drawbacks. For example, a key requirement for the superior efficiency of ML methods based on full information is that the specification of the true model should be correct and specifically concerning the likelihood function (Johnston and Dinardo, 1997) noted that "If we maximize the wrong function, we get biased estimates." (Johnston & Dinardo, 1997 p. 427). Moreover, for the not unlikely case of a (partial) misspecified model, especially in the social and behavioral sciences, effects of misspecification can spread over the estimates of the model parameters (Kumar and Dillon, 1987). The almost epistemological question about the 'truth' of models in general and especially models in social and behavioral science is treated very thoroughly by Stachowiak (1973) in his general model theory (see also Heine, 2020). According to this, models as such, and just also psychometric models, are essentially characterized by their imaging feature, the shortening feature, and their pragmatic feature (Stachowiak, 1973 pp. 131-133). Thus, in the social and behavioral sciences, according to the imaging feature and shortening feature, a true model, regardless of its complexity, is unlikely to exist and will therefore virtually always be misspecified in empirical data, certainly to varying degrees. Somewhat more pointedly, the statistician George Box (1979) already expressed this fact by stating "All models are wrong, but some are useful" (Box, 1979 S. 202). Especially the aspect of the usefulness of models, which refers to the pragmatic feature defined by Stachowiak (1973), must be the focus when using psychometric models for scaling LSA data because the declared goal here is to establish an objective scoring rule for the item responses in order to allow a fair comparison (see also Robitzsch & Lüdtke, 2021 for adesign-based perspective). The aspects briefly outlined here concerning the appropriate degree of detail and the associated extent of tolerable misspecification of psychometric models for the scaling of LSA data are closely related to the question of suitable estimation procedures for their model parameters (see, e.g., MacCallum et al., 2007). As an alternative perspective, as compared to MML (i.e., FIML) for the estimation of latent trait models with ordinal indicators (the item responses), Forero and Maydeu-Olivares (2009) suggest the use of limited information (LI) methodology for estimation (see also Bolt 2005). Such LI methodology is associated with the tradition of factor analysis (e.g., Forero & Maydeu-Olivares, 2009; McDonald, 1999), and parameter estimation, instead of assuming complete information, relies only on univariate and bivariate information in the data (Maydeu-Olivares 2001; Edwards and Orlando Edelen, 2009). Furthermore, in the LI methodology, within the concept of factor analysis, in addition to the possibility of ML estimation of the parameters, there is the alternative of ordinary least squares (OLS) estimation, which has favorable statistical properties regarding the robustness of model misspecifications. If the sampling error is neglected (by assuming an infinite sample size), the model error (as outlined above) is still very likely to be present, which represents a lack of fit of the (thus misspecified) model for the population, MacCallum et al. (2007) emphasizes that, for example, the ML estimation, in contrast to the OLS estimation, is based on the assumption that the model is exactly correct in the population and that all error is normal theory random sampling error. Put simply, the ML estimation method ignores the possible existence of a model error or the associated misspecification of the model in relation to the empirical data. In contrast, with the OLS estimate, no distributional assumptions are made, and no assumption is made about sampling error versus model error (MacCallum et al.MacCallum et al., 2007), which in turn makes OLS

likely to be more robust against a possibly misspecified scaling model. In a comparative analysis addressing the question of estimation accuracy Forero and Maydeu-Olivares (2009) show that comparable IRT model parameter estimates result from LI and ML methods. Specifically, the LI method (using OLS) provided slightly more accurate parameter estimates, and the ML method provided slightly more accurate standard errors (Forero and Maydeu-Olivares 2009). An item parameters estimation method for Rasch-type item response models, which can be attributed to the LI method is the PAIR algorithm (cf. Robitzsch, 2021a, Heine 2020). This calibration approach was introduced by Choppin(1968 see also McArthur & Wright 1985) as a sample-free calibration method for item banks in large-scale assessments, within the context of early approaches tocomparativeeducation. Choppin's row averaging approach (RA) is based on pairwise information. It has the advantage of enabling a non-iterative identification of item parameters for the Rasch model and the PCM (Choppin, 1968; Heine and Tarnai, 2015). Moreover, the pairwise RA method, as with other LI methods like pairwise conditional maximum likelihood (PCML) or the minimum chi-square method (MINCHI), reduces the computational demand for item parameter identification based on large LSA data sets (Robitzsch, 2021a). Compared to PCML, the RA approach within the LI methodology provides OLS estimators for the item parameters (cf. Mosteller, 1951b; Mosteller, 1951a; Mosteller, 1951c; Heine, 2020). As a result of a systematic comparison of several LI estimation approaches against other methods for the Rasch model, (Robitzsch, 2021a) concludes that RA and similar LI methods can be beneficial in applied research. This benefit for applied research is based on the experience from the systematic comparison of the estimation methods that RA and similar LI methods can result in less biased item parameter estimates than ML-based methods, given possible model misspecification and local dependencies in the empirical data (Robitzsch, 2021a see also Forero and Maydeu-Olivares, 2009).

Another area in the discussion about the evaluation methodology of cross-sectional LSA data relates to the aspect of the longitudinal linking of different rounds of the assessment (e.g., Robitzsch & Lüdtke, 2019; Oliveri & von Davier, 2011; Fischer et al., 2019; Gebhardt & Adams, 2007). Specifically, the principle of a successive chain linking approach, as used in the PISA rounds up to 2012, was, for example, criticized by Oliveri and von Davier (2011, 2014). They, in turn, argued for a concurrent calibration approach, including all data from the previous PISA rounds, respectively. Such an approach was applied, for example, by von Davier et al. (2019) for historical PISA data and was first introduced in the official PISA evaluation from round 2015 on wards (OECD, 2017, 2021a). Researchers von Davier et al. (2019) conclude from their study that changing the linking method had an impact on the country mean results but not on the ranking of the cross-sectional country means. Their analyses showed that the Spearman rank correlations for the mathematics competency area were $r_s = 0.994$ for the respective cross-sectional country means across all analyzed PISA rounds, a finding that von Davier et al. (2019) view as an indication of a valid or method-invariant country comparison in PISA. However, such an invariant cross-sectional rank order may not be sufficient for evaluating trend estimates. Trend estimates for a country are typically interpreted if they exceed statistical significance. However, if the choice of an analysis method impacts a country's mean of 1 or 2 points on the PISA scale, it might be consequential for the interpretation

of trend estimates. Furthermore, it could turn a statistically non-significant into a significant trend estimate, which, in turn, gets policy attention.

Although Fischer et al. (2019) found little differences among different linking methods and anchoring designs also for a longitudinal linking of competence tests in large-scale assessments scaled with the Rasch model, Robitzsch and Lüdtke (2019) demonstrated that the interpretation of national trend estimates could change when different approaches for linking and procedures to calculate standard errors are applied. In addition, Gebhardt and Adams (2007) emphasize the importance of the influence of item calibration based on different samples, that is, calibrating the items separately for each country as compared to the linear transformation approach, which uses a common set of item parameter estimates for all countries. Specifically, Gebhardt and Adams (2007) showed that the use of conditional rather than marginal means as a linking approach results in some differing conclusions regarding trends at both the country and within-country level.

Connected to the question of an appropriate linking approach for longitudinal comparisons is the question of sampling or selecting of subjects and items on which calibration and scaling are based. The question of the appropriate calibration sample and its effects on the competence measurement is proving to be increasingly relevant, especially against the increasing expansion of LSAs to other populations or states and economies (e.g., Rutkowski et al., 2019; Rutkowski & Rutkowski, 2021; Rutkowski et al., 2018). The relevance of selecting an appropriate calibration sample results from the typical fact that, for example, the PISA measuring instruments were originally developed for OECD member countries and are now increasingly used for surveys in emerging and developing countries (Rutkowski et al., 2019; Rutkowski and Rutkowski, 2021). It typically shows that around half of the existing PISA items are too difficult for these new PISA participants (Rutkowski et al., 2018), which means that an appropriate measurement of competence in low-performing educational systems can be subject to possible distortions (Rutkowski and Rutkowski, 2021). From a technical perspective, such distortion results from floor effects in the item responses (e.g., Rutkowski et al., 2019), which ultimately represent sub-optimal test targeting for certain populations, resulting from item calibration based on a more competent population than the target population.

The choice of items is an important factor in cross- and longitudinal-country comparisons insofar as it has a significant impact on the standard errors of the estimates of competence (Glas and Jehangir, 2013; Robitzsch, 2021c; Robitzsch and Lüdtke, 2019). Generalizability theory (c.f. Brennan, 2001) defines several facets for which generalization of results appears necessary. If tests are to generalize not only to the specific set of items used in the test, but to a potential universe of items in a performance domain, the source of variation in item selection (i.e., item sampling) must be taken into account. In educational research and large-scale tests, the idea of viewing the single items in a test as a realized subset of an ultimately infinite universe of items is a concept that was already introduced early on (e.g., Husek & Sirotnik, 1967, Lord & Novick, 1968). Regarding longitudinal analyses, Hutchison (2008), as well as Michaelides (2010) point out that the choice of link items between multiple studies in longitudinal analyses should preserve the interpretation of an item sampling. If the number of link items is too small or the specific choice of link items is not representative of the entire item set, biased estimates

of performance trends may result (Mazzeo and von Davier, 2008; van den Heuvel-Panhuizen et al., 2009).

Based on this exemplary and possibly not extensive presentation of some selected findings from the methodological literature on LSAs, it can be stated, at least in summary, that different methodological approaches might lead to slightly different population estimates. This phenomenon can be described as method variance.

As already described above, official PISA reporting, particularly for reporting trends in country means, is based on a variance component model to construct composite standard errors to reflect the overall uncertainty in measurement. It must be noted that this approach of composite standard errors constructed for specific comparisons of statistics, such as country means, does not necessarily follow the classical definition of the standard error as a single unique measure of dispersion $\hat{\vartheta}$ for a single estimation function $\hat{\theta}$ for a parameter θ estimated for the population. Rather, different sources of variance $(\hat{\vartheta}_1, \hat{\vartheta}_2, ...)$ are assumed, which are summed to derive the final (constructed) standard error in order to quantify the overall uncertainty of measurement in the PISA LSAs. Against the background of such a model, however, the question immediately arises as to which are the relevant variance components in a typical LSA setting.

Specifically, Robitzsch et al. (2011) argue that in a concept of generalizability, (at least) three facets in testing play an important role: The sampling or selection of subjects, the sampling or selection of items, and the choice of statistical models. The empirical findings by Robitzsch et al. (2011) indicate that the sources of variation in item sampling and model choice, which are usually neglected in publications as compared to the sampling of respondents, are not negligible. More recently, concerning item selection for linking different PISA rounds, Robitzsch and Lüdtke (2019) conclude from results of simulation as well as reanalyzing trend estimates for reading from PISA 2006 to PISA 2009 that the PISA method underestimates the standard error of the original trend estimate. Thus, the number of countries with significant trend changes decreased from 13 to 5 when using a newly proposed method for determining standard errors compared to the official PISA standard error (Robitzsch and Lüdtke, 2019).

Despite the extensive evidence on single aspects of the methodology of the official PISA data analysis, excerpts of which are reported here, there is, to our knowledge, no comparative study that shows the relative importance of these single analytic choices with respect to the error component against each other. In this study, we will add another source of variance to the standard errors constructed within the framework of a variance component model for the measurement error.

On the one hand, such a comparative analysis is interesting and important from a methodological perspective, as it can contribute to placing the relevance of single analytical decisions concerning future PISA data evaluations on an empirical basis. From a practical perspective, the findings from the reanalysis of the PISA data, taking into account the key factors of methodological variance identified here, can help to make a more realistic classification of the significance of small country mean differences, both in a cross-sectional and longitudinal comparison.

To quantitatively formulate this additional variance component resulting from the analytical decisions on methods in evaluating PISA results, we will follow two strategies. One strategy borrows from the principle used in the official PISA reporting of

Table 1 Overview for cross-sectional OECD PISA data files and resulting total database of scalable cases for mathematics

PISA round	participating countries		Number of cases in files		math scalable students		
	OECD	non OECD	SCR[a]	SDQ[a]	OECD	non OECD	Total
2003	30	11	276 165	276 165	220 670	49 995	270 665
2006	30	27	398 750	398 750	251 278	147 472	398 750
2009	34	40	515 958	515 958	298 454	217 504	515 958
2012	34	31	485 490[d]	480 174	295 416	184 758	480 174
Number of cases that can be used for scaling:					1 065 818	599 729	1 665 547

[a]SCR = *scored cognitive item response data*; SDQ = *student questionnaire data*;

[d]Includes additional regions from USA: Florida, Connecticut, Massachusetts

calculating and using composite standard errors when looking at trends, taking into account a linking error as an additional error component (cf. Eq. 4). Second, for the definition of an extended confidence interval for the country means from the PISA results, we will adopt a strategy proposed by Leamer and Leonard (1983) for evaluating the maximum upper and lower bounds of estimates from different regression models (see also Mcaleer et al., 1985; Leamer, 1985). In the subsequent method section, we will present these two approaches in more detail.

Methods

Data: compiling the historical PISA database

We use the publicly available data from the OECD download pages as the source to build up the database for the present study (see OECD 2021b). The single data sets for the PISA rounds from 2003 up to 2012 are provided by the OECD as generic text files in ASCII format. Thus, in the first processing step, the corresponding official syntax-control files (SPSS version) were used to create single SPSS data sets according to the instructions given at OECD (2020). Typically, the scored cognitive item response data (SCR) and the data from the background questionnaires, together with the plausible values (PVs) resulting from the PISA round-specific scaling for the three PISA competence areas – student questionnaire data (SDQ) – are provided as separate files per PISA round. In a second step, these separate data sets were combined for each assessment round. For this procedure, it is noticeable for some PISA rounds that the SCR and SDQ files contain a different number of cases (see Table 1). Such differences can be traced back to various obvious reasons, such as additional groups of students (over-) sampled by some countries for supplementary national research questions in some PISA rounds. Table 1 gives an overview of the number of cases contained in the SCR and SQD files together with the assumed reasons for the differences, the number of participating (OECD-) countries, and the number of scalable PISA students for the math domain in total and per cycle, which were used in the official reporting.

In order to be able to merge the single data sets into a complete data set for the further analysis steps, the variable names and their category codes, which vary from round to round, had to be harmonized. To give an example of this necessary procedure, we refer here to the question stem "Population Pyramids" (number 155) used in all PISA rounds from 2003 to 2012 with a total of four related questions (Q1 to Q4). Across the four PISA

rounds considered in our analyses, the item descriptor for question number 1 related to this question stem changes, for example, from 'M155Q01' (rounds 2003, 2006, and 2009) to 'PM155Q01' (round 2012). Even more, the category coding handled differently in each round had to be adjusted. Thus, for example, the code assigned to "not-reached" items varied from "not-reached" coded with "r" (2000, 2003) to "not-reached" coded as "8" (2006, 2009, 2012). Similarly, the categorization "missing by design" was coded "n" or "N/A" in round 2003, depending on the item (the latter was used if it was an item first introduced in round 2003), or was coded "7" across all items in round 2006, and either "7" or "N/A" in rounds 2009 and 2012, depending on the item and its first introduction into the PISA assessment respectively. In addition to this necessary harmonization of category coding, rescaling must also adopt a consistent approach to interpretation and, in turn, coding of unanswered cognitive items that had been presented to students during testing. In addition to design-related missing values, in PISA, a distinction is made within student-related missing responses between 'simple' item omissions on the one hand and so-called "not-reached" items on the other hand. The definition for the latter is based on a sequence-oriented view of item responses, based on the assumption of linearly progressive processing of the test items by the students. The typical definition for not-reached items from the technical documentation for this reads as follows: "all consecutive missing values clustered at the end of test session [...] except for the first value of the missing series, which is coded as item-level nonresponse" (OECD, 2014 p. 399, 233). However, this assumption of a linear progression can hardly be ensured empirically for the PISA rounds up to 2012 since the test items were given in paper-based form for all participating countries up to and including 2012. A consequence is that there was no reliable control over the order in which the students processed the single test items during the assessment. In addition to this uncertainty regarding the true processing order of the items by the students, some authors (e.g., Rose et al., 2010, Pohl et al., 2014) criticize that in the past PISA rounds up to 2012, the treatment of missing responses was handled differently for item calibration and proficiency scaling steps. According to this, not-reached items would be considered not administered during the item calibration based on the international calibration sample and, on the other hand, counted as wrong answers in the competency scaling. Although this different interpretation of not-reached items in the same database at different analysis steps, at least based on the official PISA documentation (c.f. OECD et al., 2002, 2009; OECD, 2005, 2012, 2014), can only be proven without a doubt for the first PISA round in 2000[1], we chose the same scoring rule for not-reached items for the present reanalysis. Moreover, Robitzsch (2020) shows that based on the definition of ability accepted for PISA in a domain of competence as the extent to which items of a given set of items in a given maximum test test-taking time, not ignoring missing responses – that is treating them as incorrect answers – leads to a

[1] a careful review of the official technical reports (c.f. OECD et al., 2002, 2009; OECD, 2005, 2012, 2014) on PISA shows that different interpretation of not-reached items for calibration and scaling is only documented beyond doubt for the PISA 2000 round. Only in the respective technical documentation for PISA 2000, it is explicitly stated that the not-reached items are handled differently for item calibration and competence scaling: "[...] using the international calibration sample of 13500 students, and not-reached items in the estimation were treated as not administered." (see OECD et al. 2002 p. 161) and further down in the publication: "[...] in this analysis [proficiency scaling], not-reached items were regarded as incorrect responses, in accordance with the computation of student ability scores [...]" (see OECD et al., 2002, p. 161, Additions in square brackets).

more valid definition of the proficiency itself (Robitzsch and Lüdtke, 2021; Robitzsch, 2021b). Based on this reasoning, we chose to score all student item omission and not-reached items as wrong answers in both, item calibration and proficiency scaling – (but also Pohl et al., 2021see for an alternative view).

As described in the introduction in the section on international scaling principles, in single PISA rounds, specific items were excluded from the official reporting for some countries ("dodgy" items, see, e.g. OECD, 2005, p. 122). For the domain of mathematics, this concerned nine items on the national level and one item on the international level (OECD, 2005, p. 190) in the 2003 PISA round, seven items on the national level and none on the international level (OECD et al. 2009, p. 216) in 2006, eleven items on the national level and one item on the international level (OECD, 2012, p. 196) in 2009, and finally six items on the national level and one item on the international level (OECD, 2014, p. 232) in 2012. The technical documentations for each PISA round note that for all national item deletions "All deleted items were recoded as not applicable" (OECD, 2005, p. 190) (OECD et al., 2009, p. 216) (OECD, 2012, p. 195) (OECD, 2014, p. 231), which typically corresponds to the code "7", which was interpreted as missing in our reanalyses, thus excluding the item in question from our analyses (for the respective country), as was the case with international reporting. Items deleted at the international level are not included in the downloadable database on the OECD pages.

Analytical decisions and methods applied in the present study

Four key methodological factors are considered in the present reanalysis. These key factors relate to the estimation method used for item calibration (*estimation*), the selection of the included (link) items (*items*), the selection of the calibration sample (*sample*), and the calibration principle in linking the single PISA rounds (*calibration*).

With respect to the basic linking of the PISA rounds, we first distinguish between a concurrent calibration (concurrent) of the items using the combined database comprising all PISA rounds and a separate item calibration for each PISA round with a subsequent chain-linking approach (chain) in order to quantify the influence of this analytical decision (factor *calibration*) on the PISA rankings and trend statements between and within countries (see Robitzsch & Lüdtke, 2019,Gebhardt & Adams, 2007, von Davier et al., 2019). In addition, we alter the country samples to rest the item calibration on (factor *sample*). Specifically, we distinguish four factor levels. First, "OECD 2003": That is OECD countries only, as defined for the survey round in 2003 (possibly reduced in later rounds if a country did not participate in the respective round). Second, "OECD by CYC": that means OECD countries only, as defined for the respective survey round. Third, "ALL 2003" which includes a constant number of all participating countries from the (first) survey round in 2003 (possibly reduced in later rounds if a country did not participate in the respective round). Finally, "ALL" includes all available data resulting in a growing database of participating countries across all survey rounds (see Table 1). Concerning the selection of the (link) items used, we distinguish two factor levels (factor *items*). On the one hand, I mathematics items available in total across all four PISA rounds (2003 - 2012, $I = 158$) are included in the calibration step (factor level: total), or on the other hand, only those items that were used completely across all four rounds (from 2003 to 2012, $I = 34$; factor level:

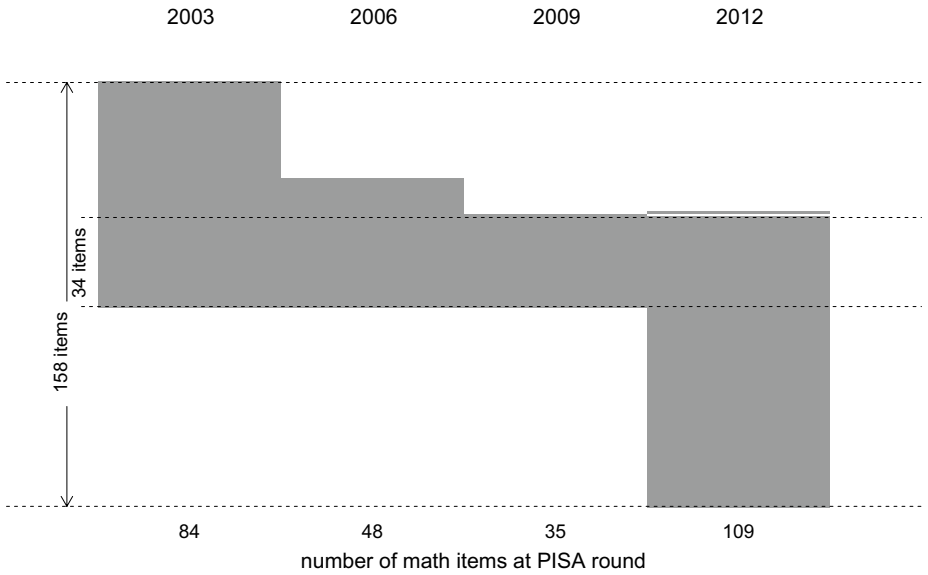

Fig. 1 PISA mathematics item linking across four rounds; $I = 158$; Common mathematics items linking single rounds: $I = 48$ (2003-2006); $I = 35$ (2006-2009); $I = 34$ (2009-2012); $I = 35$ (2003-2012)

complete). In the period from 2003 to 2012, we considered and reanalyzed in the present study, $I = 158$ mathematics items were used. Of these $I = 158$ items, $I = 84$ were used in the 2003 round, $I = 48$ in the 2006 round, $I = 35$ in the 2009 round, and finally, $I = 109$ in the 2012 round. If we look at the overlaps of two adjacent PISA rounds with regard to the common math items used, we find that $I = 48$ items connect the rounds 2003 and 2006, $I = 35$ items connect the rounds 2006 and 2009, and $I = 34$ items connect the rounds 2009 and 2012 (see Fig. 1). Figure 1 provides a visualization of this item linking across the four rounds of PISA, while items existing in the respective PISA round are marked in grey. Since Fig. 1 only gives a visual impression of item linking across the four PISA rounds covered here, an excel file exists as an electronic supplement for detailed documentation (including item names) of the item linking at the OSF repository (https://osf.io/ubvaq/?view_only=0aaff3adfbd24f1488b81e4508c8d9e8).

The concurrent and the separate item calibrations are subject to the assumption of the partial credit model (PCM Masters, 1982) as being the scaling model PISA mathematics items. In order to capture the influence of different methods of parameter estimation during calibration and scaling, we consider two different parameter estimation methods in the reanalysis of the PISA data, as implemented in two *R* packages (factor *estimation*). First, we use the *R* package `TAM` (see Robitzsch et al., 2021) to apply an MML-based estimation principle. Next to different approaches in parameter estimation for one- and multidimensional item response models, the *R* package TAM implements the multidimensional random coefficients multinomial logit model (Adams et al., 1997) and it supports latent regression models as well as plausible value imputation. In this respect, `TAM` can be viewed as an open-source *R* version of the IRT software ConQuest (Wu et al., 2012), which was officially used to calibrate and scale the PISA data from 2000 up to 2012. Second, we use the *R* package `pairwise` (see

Heine, 2021) to realize a pairwise non-iterative row averaging approach (PAIR) for model estimation, according to Choppin (1968). The resulting item parameters can be viewed as least square estimators (Mosteller, 1951b, a, c), which, as already discussed above, can have advantageous properties compared to ML estimators, especially against the background of possible model misspecifications. Moreover, unlike likelihood-based methods, the PAIR approach does not involve iterative estimation (e.g., cf. Heine, 2020). The PAIR procedure relies on the (conditional) pairwise comparisons of the item category frequencies and makes use of the bivariate information of the item associations. Therefore, the resulting item parameters can also be regarded as LI estimates (Bolt, 2005; Christoffersson, 1975; Forero and Maydeu-Olivares, 2009; Lance et al., 1988; Maydeu-Olivares, 2001).

Analogous to the official evaluation procedure for PISA 2000 to 2012, we also apply a multistage calibration and scaling process in this reanalysis. All analyses were performed in the free statistics environment *R* (R Core Team, 2022 ,version 4.2.0) and are documented in single *R* scripts (see Additional file 1). In the first step, the items included are calibrated (using both estimation principles as described above). Then their parameters are assumed to be fixed when scaling or drawing the PVs as competence estimates. The PVs were drawn by calling the function `tam.pv()` from the *R* package `TAM` with the corresponding defaults for the argument defining the number of PVs to be drawn, resulting in a drawing of 10 plausible values (`nplausible = 10`). In the case of concurrent item calibration, the fixed item parameters refer in the same way to all PISA rounds. In the case of separate calibration, the item parameters from the respective PISA round are used as a basis when drawing the PVs. As for the official OECD reporting, we use the variance of the item parameter estimates from the individual rounds to quantify the linking error in the subsequent linking of the different rounds (cf. Eqs. 3 and 4).

The resulting mathematics competency estimates on the logit metric are subjected to a transformation in the final step so that they lie on the PISA metric ($M = 500$, $SD = 100$), in a way that the mean estimates are anchored to the mean estimates from the PISA round 2003. Specifically, the anchoring of the results from the rescaling is achieved by a linear transformation of the logit scale to match the typical PISA reporting scale:

$$\theta_{\text{PISA}} = a\theta_{logit} + b \tag{5}$$

The rescaling steps described above are performed according to the four factors (*items*, *sample*, *calibration* and *estimation*) with their respective factor levels for $2 \times 4 \times 2 \times 2 = 32$ analytical scenarios (see e.g., Table 3). The resulting matrix with PISA country mean estimates and their respective errors are then subject to further analysis to quantify the relative influence of the different analytic choices on the two central outcome measures, namely country ranking and development trends for PISA Mathematics proficiency scores between and within countries. For these analyses, variance component analyses are performed on the matrix with country mean estimates using the *R* package `lme4` (Bates et al., 2015, 2021). These analyses thus show the relative influence of single analytic choices in item calibration and competence scaling on three trend analyses comparing 2003 to 2006, 2006 to 2009, and 2009 to 2012 rounds of

PISA on the development of mathematical competence (see Table 4), as well as for the cross-sectional scaling approach (see Table 5).

To vividly illustrate the cumulative effects of the method variance from the analytic choices and the resulting widened confidence intervals for the point estimates of country mathematics performance, we draw on the results of three groups of countries representing three broad areas of the international distribution of PISA mathematics performance. For the highly proficient group, we select the countries Hong Kong - China (HKG), South Korea (KOR), and Finland (FIN). The group of countries with intermediate mathematics literacy is represented by the countries Belgium (BEL), Australia (AUS), Germany (DEU), Iceland (ISL), and Denmark (DNK). And finally, the group with medium to rather low mathematical literacy in the PISA mathematics ranking is represented by the countries Ireland (IRL), United States of America (USA), Portugal (PRT), Spain (ESP), and Italy (ITA)—see Fig. 2 and Table 2. For these countries, upper and lower bounds of confidence intervals (BI_{upper} and BI_{lower}) for the suspected true country means as well as pooled compound standard errors ($SE_{comp.}$) for the respective aggregated country mean across $M = 32$ means from the rescaling are calculated to account for the method variance component referring to different analytical choices. Specifically, the upper and lower bounds of confidence intervals are constructed according to the principle shown in Eqs. (6) and (7) by considering the maximum confidence interval that would be obtained when joining all intervals.

$$BI_{upper} = \max_{m=1,\ldots,M} \left\{ \hat{\mu}_m + 1.96 \cdot SE(\hat{\mu}_m) \right\} \tag{6}$$

$$BI_{lower} = \min_{m=1,\ldots,M} \left\{ \hat{\mu}_m - 1.96 \cdot SE(\hat{\mu}_m) \right\} \tag{7}$$

The respective (error) range (Δ_{BI}) of the (theoretical) true country mean is then defined as the difference between the upper and lower bounds ($\Delta_{BI} = BI_{upper} - BI_{lower}$); see Fig. 2 and column Δ_{BI} in Table 2) .

Note that this approach of defining the upper and lower bounds for a conservative confidence interval based on extreme values $\hat{\mu}_{max}$ and $\hat{\mu}_{min}$ from the country mean estimates and their respective standard errors across 32 analytical decisions follows a principle proposed by Leamer and Leonard (1983). Leamer and Leonard (1983) propagated this approach to represent the uncertainty in estimates from regression models that arise from ambiguity in the choice of a model (see also Mcaleer et al.,1985; Leamer, 1985 for a more detailed discussion of this approach).

To additionally calculate a method variance-supplemented compound standard error ($SE_{comp.}$) for the aggregated country means across the $M = 32$ single means from the combinations of analytic decisions, the square root of the variance of the single means is added by the pooled standard errors of the $M = 32$ single estimates (see Eq. 8 and column $SE_{comp.}$ in Table 2).

$$SE_{comp} = \sqrt{\frac{1}{M}\sum_{i=1}^{M} SE(\hat{\mu}_m)^2 + \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\mu}_i - \frac{1}{M}\sum_{i=1}^{M}\hat{\mu}_i\right)^2} \tag{8}$$

Table 2 Comparison of Country Means Standard Errors and Confidence Intervals for OECD Method and Rescaling for 13 Countries

country	2003					2006				
	OECD Meth.		Rescaling			OECD Meth.		Rescaling		
	M	*SE*	$M_{aver.}$ [a]	$SE_{comp.}$ [b]	Δ_{BI} [c]	*M*	*SE*	$M_{aver.}$ [a]	$SE_{comp.}$ [b]	Δ_{BI} [c]
HKG	548.30	(3.75)	546.32	(5.46)	11.46	549.61	(2.13)	551.78	(3.37)	8.34
KOR	540.15	(2.69)	539.41	(3.45)	7.80	548.98	(2.92)	549.92	(3.97)	9.01
FIN	542.26	(1.77)	543.33	(2.46)	5.81	549.55	(2.10)	551.51	(3.51)	8.71
BEL	530.43	(2.05)	529.34	(2.77)	6.51	525.53	(2.34)	525.25	(3.04)	7.18
AUS	523.60	(1.90)	522.85	(2.62)	6.05	521.05	(1.80)	522.30	(2.89)	6.76
DEU	503.48	(2.95)	504.51	(3.91)	8.92	508.84	(3.07)	510.46	(4.80)	10.89
ISL	514.48	(1.57)	512.60	(3.60)	8.10	508.24	(1.91)	507.14	(3.31)	7.99
DNK	513.58	(2.52)	512.16	(3.85)	8.29	515.34	(2.30)	514.23	(3.35)	8.50
IRL	501.83	(2.15)	500.63	(3.37)	7.40	504.40	(2.34)	503.93	(3.28)	7.59
USA	482.53	(2.59)	481.68	(3.79)	8.36	478.05	(3.24)	476.27	(5.00)	11.44
PRT	466.87	(2.99)	466.36	(3.33)	7.37	471.24	(2.33)	470.93	(3.00)	7.90
ESP	485.00	(2.20)	484.66	(2.63)	6.03	483.79	(1.83)	483.08	(2.55)	6.49
ITA	466.53	(2.53)	470.06	(6.21)	13.17	466.96	(1.81)	467.41	(2.42)	6.25
country	**2009**					**2012**				
	OECD Meth.		**Rescaling**			**OECD Meth.**		**Rescaling**		
	M	*SE*	$M_{aver.}$ [a]	$SE_{comp.}$ [b]	Δ_{BI} [c]	*M*	*SE*	$M_{aver.}$ [a]	$SE_{comp.}$ [b]	Δ_{BI} [c]
HKG	556.29	(2.48)	558.04	(3.52)	9.43	566.40	(2.83)	564.73	(5.54)	14.65
KOR	548.65	(2.92)	550.83	(4.07)	9.53	557.53	(3.93)	554.87	(6.97)	16.56
FIN	541.87	(1.71)	543.48	(2.74)	6.80	523.34	(1.55)	527.71	(5.51)	14.51
BEL	519.56	(1.80)	521.03	(2.65)	6.89	519.83	(1.85)	518.59	(3.79)	10.07
AUS	517.65	(2.06)	519.05	(2.74)	6.86	509.03	(1.43)	511.85	(3.89)	11.20
DEU	516.03	(2.40)	516.58	(3.15)	7.59	518.02	(2.48)	519.79	(3.92)	10.12
ISL	510.46	(1.75)	510.92	(2.62)	7.00	496.91	(1.73)	497.21	(3.16)	9.74
DNK	505.41	(2.48)	507.05	(3.06)	7.46	504.78	(2.06)	504.53	(3.41)	9.31
IRL	493.36	(2.01)	493.63	(3.08)	8.42	505.98	(1.92)	507.35	(3.16)	9.05
USA	490.05	(2.80)	490.01	(3.95)	9.77	485.62	(3.05)	485.98	(4.06)	10.99
PRT	489.33	(2.21)	489.25	(3.76)	10.71	492.58	(3.31)	491.52	(5.16)	13.34
ESP	487.68	(1.85)	487.76	(2.88)	7.54	489.83	(1.63)	490.58	(2.58)	8.09
ITA	486.19	(1.48)	486.40	(3.00)	7.90	490.18	(1.75)	489.61	(3.19)	8.96

[a] Mean aggregated Country mean ($M_{aver.}$) across 32 combinations of analytical decisions

[b] compound Standard error ($SE_{comp.}$) accounting for method variance across 32 combinations of analytical decisions

[c] Range of error from constructed confidence interval (Δ_{BI}) based on highest and lowest country mean estimates and their respective standard errors to account for methodvariance.

Note that both approaches of either constructing upper and lower bounds of confidence intervals or compound standard errors for the estimates across the analytic decisions are, in pure principle, subject to the same rationale, but nevertheless carry slightly different meanings in terms of substance. While the upper and lower bounds of the confidence interval (BI_{upper} and BI_{lower}) provide a conservative definition for the error range of the (theoretical) true country mean, taking into account different methodological approaches (models), $SE_{comp.}$ represents the standard estimation error, enriched by method variance, for the mean country performance estimator aggregated across the 32 analytical decisions.

Results

Figure 2 gives a first visual impression of the impact of the additional variance components contributing to widening the confidence intervals (Δ_{BI}) for the country mean estimates of mathematics proficiency. Thus, the area between the error bars in solid lines represents the sum of the method variance as it results from considering the four different key analytical decisions. The error bars in dashed lines represent the range of the confidence intervals based on the means and standard errors resulting from applying the method used in the original OECD reporting.

The respective country means formed over the 32 combinations of analytical decisions as well as the respective compound standard errors ($SE_{comp.}$) of estimation as well as the confidence range around the mean (Δ_{BI}), which both include the method variance across the $M = 32$ combinations of analytical decisions, are shown in Table 2 for the 13 countries reported here as examples.

Overall, Fig. 2 shows that the range of confidence intervals of the mean estimates increases when different analytical choices for rescaling are taken into account for all countries. Specifically, the comparison of the two error ranges in Fig. 2 from the rescaling on the one hand and the outcome based on the method used in the OECD reporting on the other hand for the country means suggest that for some of the country comparisons, different conclusions must be drawn from this with regard to the meaning of the country differences. While the confidence intervals resulting from applying the method from the original OECD reporting in 2003 do not overlap for the mean comparison between Germany (DEU) vs. Island (ISL), as well as in 2006 between Germany (DEU) vs. Denmark (DNK), the respective confidence intervals resulting from the rescaling do overlap. The same is true regarding the comparison between Ireland (IRL) vs. Italy (ITA) in cycle 2009, Hong Kong (HKG) vs. Korea (KOR) in cycles 2009 and 2012, as well as Australia (AUS) vs. Belgium (BEL) and Germany (DEU) within cycle 2012, USA vs. Island (ISL) and Spain (ESP) vs. Island (ISL) in cycle 2012 (see Fig. 2).

Table 2 shows the aggregated country mean estimates of mathematical proficiency of the 13 countries as they result from the 32 combinations of analytical decisions. The associated errors ($SE_{comp.}$) include the method variance resulting from the analytical decisions and the uncertainty of the respective aggregated point estimates. In addition, the range of the constructed confidence intervals (Δ_{BI}) provides a conservative estimate for the confidence range for the suspected true country means (cf. columns $SE_{comp.}$ or Δ_{BI} under the heading Rescaling in Table 2, respectively).

Overall, the comparison of the estimates based on the method of the OECD reporting with the rescaling estimates shows that the 32 combinations of analytical decisions have a substantial impact on the country means and the associated errors and confidence intervals.

The respective absolute differences between the country means based on the method of the OECD reporting and the rescaling range from a minimum of 0.04 PISA points (USA, 2009) to a maximum of 4.4 PISA points (Finland, 2012). In these two cases, as well as for the vast majority of comparisons between the estimates based on the method used in the OECD reporting and rescaling, the rescaling estimates tend to lead to a higher mathematics proficiency mean estimate of the respective country (see Table 2). The comparison of pooled compound standard error ($SE_{comp.}$), accounting for the

methodological variance from the 32 combinations of analytical choices, with the standard error (*SE*) based on the method of the OECD reporting, shows the general tendency for the standard errors to increase on average by an amount equal to a third of the initial standard errors. Thus, the average error using the method of the OECD reporting for the 13 countries is $SE = 2.36$, the average standard error from rescaling, including the method variance is $SE_{comp.} = 3.73$ (see also Table 2). To get an impression of the influence of the proportion of error related to the different analytical decisions, we calculated the error ratio (ER), which is defined as the quotient of the averaged standard error (across the 32 rescaling models) and the respective compound standard error ($ER = SE_{comp.}/SE_{aver.}$; see Robitzsch, 2022b; Buckland et al., 1997). The mean error ratio calculated in this way across the three groups of countries considered here as an example is $ER = 2.04$ for the 2012 PISA round, $ER = 1.46$ for 2009, $ER = 1.49$ for 2006, and finally $ER = 1.57$ for 2003; (see also Table 3 for the German sub-sample).

Table 3 presents the different mean estimates for the German PISA sub-sample as the result of each of the 32 combinations of analytical decisions. An examination of the single mean estimates shows a clear influence of the item selection on the resulting mean estimates over the PISA rounds of 2003 and 2006. For example, the mean estimates for the 2003 round using the total set of mathematics items (factor level *total*) across all other analytic decisions are distributed narrowly around a value of 503 scale points on the PISA metric, whereas the mean estimates based only on the set of complete link items (factor level *complete*) fluctuate around a value of 505 while a comparable picture emerges for the 2006 PISA round, the effect of item selection seems to be less pronounced for the 2009 and 2012 rounds (see Table 3). Note that the single results for all other countries as presented in Table 3 for the German sub-sample are provided as a 'csv' file in the OSF repository at https://osf.io/ubvaq/?view_only=99ba91912c0f450eaa198be51e88ded9.

In contrast to the results reported so far, which refer to 13 selected countries, which were selected here to be representative of the international distribution of competence in mathematics, the variance components analyses refer to all countries available for trend analyses in the corresponding rounds. The comparative analysis of the variance components from the four analytical decisions shows for the trend analyses that the factor (*items*) of item selection covers the relatively largest proportion of the variance. In addition to the comparison of the main effects of the four factors (*items*, *sample*, *calibration* and *estimation*), the observation of the first-order interaction effects indicates in a comparable way that the interaction term *country:item* covers the relatively largest share of the variance (see Table 4). This effect of item selection is similar, but not identical, in the cross-sectional analysis. In addition to a large country effect as expected, it can be seen that the factor of different item selection (*items*) covers an increasing proportion of the variance as the number of available items increases throughout the PISA rounds. Over the four PISA rounds, the respective decreasing relative share of variance from the different types of item calibration (factor *calibration*) is in the opposite direction (see Table 5). For the second-order interaction effects, both the cross-sectional and the trend analysis show that the interaction term *item:country* is the most pronounced (see row 8 in Table 4 and row 7 in 5). For the three-way interactions, no systematic effects can be found across the three trend estimates or four cross-sectional analyses.

Table 3 Combination of analytical decisions and PISA Mathematics Outcome for Germany

	Items	Sample	Calibration	Estimation	2003		2006		2009		2012	
					M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
1	Total	ALL	Chain	PAIR	503.1	(2.9)	508.7	(3.0)	518.0	(2.4)	516.6	(2.4)
2	Total	ALL	Chain	MML	503.7	(3.0)	508.1	(3.1)	515.9	(2.5)	515.9	(2.4)
3	Total	ALL	Concurrent	PAIR	503.6	(2.8)	508.7	(3.1)	517.4	(2.5)	517.9	(2.5)
4	Total	ALL	Concurrent	MML	503.1	(2.9)	509.2	(3.1)	517.5	(2.4)	517.4	(2.6)
5	Total	ALL 2003	Chain	PAIR	503.1	(2.8)	509.2	(3.2)	517.2	(2.5)	519.0	(2.6)
6	Total	ALL 2003	Chain	MML	503.2	(2.9)	508.9	(3.1)	516.2	(2.5)	518.1	(2.6)
7	Total	ALL 2003	Concurrent	PAIR	503.2	(2.9)	508.5	(3.2)	517.8	(2.3)	519.8	(2.5)
8	Total	ALL 2003	Concurrent	MML	503.7	(2.8)	508.1	(3.1)	516.4	(2.6)	518.2	(2.6)
9	Total	OECD 2003	Chain	PAIR	503.3	(2.9)	509.3	(3.2)	516.7	(2.4)	520.3	(2.5)
10	Total	OECD 2003	Chain	MML	503.4	(2.9)	508.3	(3.0)	516.5	(2.5)	518.7	(2.5)
11	Total	OECD 2003	Concurrent	PAIR	503.6	(2.8)	508.7	(3.1)	516.9	(2.5)	521.0	(2.6)
12	Total	OECD 2003	Concurrent	MML	504.4	(2.9)	508.8	(3.0)	516.1	(2.5)	519.6	(2.5)
13	Total	OECD by CYC	Chain	PAIR	503.4	(2.8)	509.0	(3.1)	516.5	(2.6)	519.2	(2.5)
14	Total	OECD by CYC	Chain	MML	503.5	(2.9)	508.8	(3.1)	516.0	(2.4)	518.0	(2.5)
15	Total	OECD by CYC	Concurrent	PAIR	503.5	(2.9)	509.0	(3.0)	517.9	(2.3)	520.3	(2.5)
16	Total	OECD by CYC	Concurrent	MML	503.8	(2.8)	508.7	(3.1)	516.8	(2.5)	520.0	(2.5)
17	Complete	ALL	Chain	PAIR	505.3	(2.6)	512.4	(2.9)	516.3	(2.7)	520.7	(2.2)
18	Complete	ALL	Chain	MML	505.3	(2.6)	512.0	(2.9)	516.6	(2.4)	520.2	(2.4)
19	Complete	ALL	Concurrent	PAIR	505.8	(2.6)	512.4	(2.9)	516.9	(2.7)	521.2	(2.2)
20	Complete	ALL	Concurrent	MML	505.3	(2.8)	512.0	(2.9)	516.8	(2.4)	521.6	(2.0)
21	Complete	ALL 2003	Chain	PAIR	505.2	(2.9)	511.5	(2.8)	515.7	(2.5)	520.1	(2.2)
22	Complete	ALL 2003	Chain	MML	504.9	(2.7)	512.5	(3.0)	516.9	(2.5)	521.4	(2.4)
23	Complete	ALL 2003	Concurrent	PAIR	505.6	(2.7)	512.3	(2.9)	516.9	(2.6)	521.5	(2.4)
24	Complete	ALL 2003	Concurrent	MML	506.3	(2.8)	512.6	(3.3)	516.4	(2.6)	520.7	(2.4)
25	Complete	OECD 2003	Chain	PAIR	505.3	(2.8)	512.8	(3.1)	516.4	(2.6)	521.3	(2.3)

Table 3 (continued)

	Items	Sample	Calibration	Estimation	2003		2006		2009		2012	
					M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
26	Complete	OECD 2003	Chain	MML	505.7	(2.7)	512.2	(2.9)	516.6	(2.6)	520.1	(2.6)
27	Complete	OECD 2003	Concurrent	PAIR	505.4	(2.6)	511.9	(3.2)	515.8	(2.3)	521.4	(2.5)
28	Complete	OECD 2003	Concurrent	MML	505.5	(2.6)	511.8	(2.9)	516.0	(2.6)	520.5	(2.2)
29	Complete	OECD by CYC	Chain	PAIR	506.2	(2.6)	512.0	(2.9)	515.9	(2.5)	519.8	(2.2)
30	Complete	OECD by CYC	Chain	MML	505.8	(2.9)	512.2	(2.9)	515.7	(2.6)	520.7	(2.3)
31	Complete	OECD by CYC	Concurrent	PAIR	505.3	(2.7)	512.0	(3.0)	516.5	(2.5)	520.5	(2.3)
32	Complete	OECD by CYC	Concurrent	MML	506.0	(2.8)	512.4	(3.1)	515.5	(2.6)	521.3	(2.4)
Summary :						(2.8)[d]		(3.0)[d]		(2.5)[d]		(2.4)[d]
					504.5[a]	(3.9)[b]	510.5[a]	(4.8)[b]	516.6[a]	(3.2)[b]	519.8[a]	(3.9)[b]
						1.40[c]		1.58[c]		1.26[c]		1.62[c]

Mean aggregated country mean (M_{aver})[a] and respective compound standard error (SE_{comp})[b] and error ratio (ER^{I}) as the quotient of the averaged standard error across 32 combinations of analytical decisions ($SE_{aver.}$[d]) and the respective compound standard error (SE_{comp})[b]).

Table 4 Standard Deviations of Different Factors of Analytical Decisions for Trend Estimates in PISA Mathematics Obtained from a Variance Component Analysis

	Source of variance	2003-2006[a]	2006-2009[b]	2009-2012[c]
1	Country	7.28	11.51	8.96
2	Sample	0.14	0.12	0.43
3	Calibration	0.00	0.36	0.47
4	Items	0.46	0.00	2.27
5	Estimation	0.00	0.00	0.00
6	Sample:country	0.00	0.00	0.00
7	Calibration:country	0.14	0.03	0.29
8	Items:country	2.59	2.04	4.10
9	Estimation:country	0.00	0.16	0.12
10	Sample:calibration	0.00	0.11	0.07
11	Sample:items	0.10	0.04	1.04
12	Estimation:sample	0.00	0.00	0.00
13	Items:calibration	0.00	0.19	0.00
14	Estimation:calibration	0.00	0.00	0.29
15	Estimation:items	0.11	0.33	0.02
16	Sample:calibration:country	0.10	0.00	0.00
17	Sample:items:country	0.25	0.20	0.20
18	Estimation:sample:country	0.00	0.00	0.00
19	Items:calibration:country	0.15	0.00	0.00
20	Estimation:calibration:country	0.09	0.08	0.23
21	Estimation:items:country	0.23	0.00	1.02
22	Sample:items:calibration	0.09	0.02	0.03
23	Estimation:sample:calibration	0.11	0.08	0.11
24	Estimation:sample:items	0.10	0.01	0.10
25	Estimation:items:calibration	0.08	0.13	0.23
26	Residual	0.55	0.54	0.58

Variance components including number of countries: [a]40, [b]57 and [c] 63.

Discussion

With the present work, it could be demonstrated that different analytical decisions in calibration and scaling can have a decisive influence on both cross-sectional country comparisons and longitudinal statements on changes in competencies as assessed in large-scale assessments (LSA). The findings from the analyses shown here based on the PISA data from the four rounds of 2003 to 2012 suggest that item selection is of particular importance in a relative comparison of the core factors of analytic decisions included here. This applies equally to cross-sectional comparisons of individual countries and to statements about substantial changes in mean estimates of proficiency within individual countries over the four survey periods. The meaningful effects of the second-order country-by-item interaction also found suggest that single items might function differently in different countries, which may subsequently impact the central outcome in the cross-sectional country means and, hence, PISA ranking charts. Based on the plausible assumption, not only supported by the methodological literature on the scaling of large-scale data, that the factors of analytical decisions considered in the analyses conducted here represent at least a substantial part of the possible method variance that exists in the evaluation of LSA data, conclusions can be drawn about the meaningfulness

Table 5 Standard Deviations of Different Factors of Analytical Decisions for Cross-Sectional Country Means in PISA Mathematics Obtained from a Variance Component Analysis

	Source of Variance	2003[a]	2006[a]	2009[a]	2012[a]
1	Country	48.69	47.23	44.52	43.30
2	Sample	0.00	0.82	1.80	0.00
3	Items	0.00	0.80	1.02	1.19
4	Estimation	0.02	0.71	0.36	0.00
5	Calibration	4.78	1.05	0.81	0.01
6	Country:sample	0.10	0.02	0.09	0.00
7	Country:items	2.58	2.02	1.71	4.63
8	Country:estimation	0.10	0.07	0.09	0.00
9	Country:calibration	0.15	0.17	0.12	0.12
10	Sample:items	0.00	0.00	0.00	0.97
11	Sample:estimation	0.00	0.42	0.38	0.16
12	Sample:calibration	0.00	0.22	0.04	0.02
13	Estimation:items	0.03	0.08	0.47	0.86
14	Calibration:items	0.00	0.00	0.08	0.36
15	Calibration:estimation	0.02	1.22	1.58	0.20
16	Country:sample:items	0.00	0.00	0.00	0.12
17	Country:sample:estimation	0.00	0.00	0.00	0.00
18	Country:sample:calibration	0.05	0.06	0.03	0.06
19	Country:estimation:items	0.09	0.00	0.10	0.60
20	Country:calibration:items	0.04	0.10	0.05	0.07
21	Country:calibration:estimation	0.06	0.00	0.09	0.11
22	Sample:estimation:items	0.00	0.00	0.03	0.15
23	Sample:calibration:items	0.00	0.24	0.22	0.33
24	Sample:calibration:estimation	0.00	3.87	6.65	3.03
25	Calibration:estimation:items	0.01	0.10	0.26	0.51
26	Residual	0.37	0.44	0.42	0.45

[a] Variance components considering 40 countries.

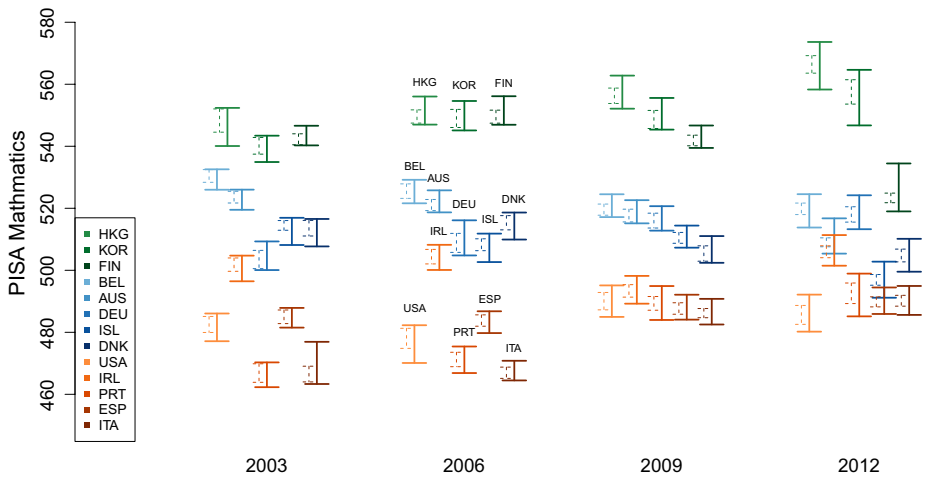

Fig. 2 PISA mathematics country means with confidence intervals constructed from methodological variance (solid lines); Method from OECD reporting with confidence intervals (dashed lines)

of country-mean differences. For example, a systematic inspection of the constructed confidence intervals in this study suggests that the confidence intervals of the inferred true country means are about twice as large as the error from the official reports, after accounting for possible differences in analytic choices. This overarching finding helps to put into relative terms the "horse race communication" (see Ertl, 2020) regarding the PISA country rankings, which has been forced not least by media reporting (see Johansson, 2016). Against the background of the findings from the present analyses, a rather cautious interpretation of such differences in the PISA rankings seems to be indicated, especially in the case of only small country mean differences. Such a more moderate view of, in particular, small country mean differences could help to put into relative terms the pure governing by numbers (see Grek, 2009) in the area of educational policy development and, at the same time, broaden the view of country-specific differences in educational prerequisites. Such an expanded view on LSA data beyond rankings could foster more promising nuanced analyses within countries (see Singer & Braun, 2018), such as on skill distribution within population groups, rather than pure mean aggregates, or supplementing and evaluating country-specific longitudinal components of LSA studies.

Our approach to rescaling the PISA mathematics data 2003 – 2012 also implies prior decisions on the analytical procedure. The analytical decisions we take into account are based on theoretical considerations or existing findings from the methodological literature on LSAs. Some of our analytical decisions may still be part of an open controversial debate. For example, the decision to treat missing responses, both omitted items, and items not-reached, as incorrect responses, as we have dealt with in our analyses, may be controversial. From the rather extensive theory and years of research on missing data comes a rich arsenal of widely varying and nontrivial methods for dealing with these missing data points depending on the nature of the data failure process (see, e.g., Rubin, 1987). However, no matter how fancy these approaches are, they are all ultimately based on more or less strong assumptions whose ultimate hold is usually difficult to prove. In our analyses, we choose the uniform interpretation of items not-reached or omitted as incorrect answers for item calibration and person scaling steps. We justify this decision with reference to the literature (cf. Robitzsch A & Lüdtke, 2021; Robitzsch, 2021b) with the sound construct definition underlying the PISA competency domains, substantiating the selection of specific items in the test that operationalize the construct. In such a design-based perspective, it is not left to the test takers to define the construct by their ad hoc selection of the items they want to work on. Finally, yet importantly, the finding that item selection typically has a considerable influence on measurement, which was also confirmed in our analyses, supports such a design-based perspective to the detriment of a model-based perspective that, for example, strives for a better model fit. The design-based view of latent variable operationalization parallels the formative latent variable model in that the constructs to be measured consist of a fixed operationalization—in contrast to the reflective latent variable model in which the single items are seen as interchangeable indicators and (in the best case) equally good and representative of a construct to be measured. Finally, we would note that we are well aware of the ongoing debate about the treatment of omitted and not-reached items in LSAs and the lack of an ideal solution for these missing item responses. However, the very magnitude and importance of this open debate is more likely to warrant a stand-alone paper, so we

would like our present analyses and findings to be understood as a stimulating starting point for further research. In this respect, based on further empirical research, we would welcome a critical discussion of the potential impact of different approaches to item nonresponse on scaling results as a further factor of analytic decisions in evaluating LSA data (see Robitzsch, 2022a). This is particularly the case as increasingly recent data from PISA and other LSAs typically reinforce the trend that the proportion of missing data varies across countries.

Another criticism could be that in this paper, we do not include the analysis methodology that has been expanded in the current PISA rounds as an additional factor in analytical decisions. However, it must be stated that our analytical decisions relate to the data analyzed here, which covers the period up to the PISA survey in 2012, and the currently expanded (new) analysis methodology used for the official OECD reporting has only been applied to the PISA data since the survey in 2015. In this respect, we take into account the methodology that was officially used in the period of the data analyzed here. The decision to only base our analyses on the data up to the 2012 round of PISA and, therefore, to limit ourselves to the corresponding methodologies was motivated by two essential aspects that can be found in the transition from PISA 2012 to 2015. On the one hand, the general assessment mode changed from paper-based to computer-based surveys and, on the other hand, the model for competence scaling changed to a 2PL model together with the change of the core contractor. These two changes have resulted in numerous methodologically oriented analyses and papers on their effects, which is why we refrained from analyzing these two aspects again for this paper to strive for clarity of the expected findings.

Conclusions

The results of the present study should not be misunderstood as an undifferentiated overall critique of the analytical decisions made in past PISA rounds for the official PISA reporting. After all, analytical decisions are always based on current methodological knowledge and its practical feasibility within large-scale comparative studies such as PISA. But this methodological knowledge is subject to constant change with constant new developments, and, above all, the practical implementation of the latest methods in LSAs is not always immediately possible or even appropriate. The central message of this paper is the finding that analytic decisions to scale LSA data like the PISA data affect the results. In this context, the three factors of choosing different calibration samples, estimation procedures, and linkage methods tend to exert little influence on the country-specific cross-sectional and trend estimates. However, the choice of different link items, which also represent the basis for the operationalization of the construct when rescaled accordingly, appears to have a decisive impact on country rankings and trends between and within countries. To summarize, different analytic choices in the evaluation of LSAs can be seen as a so far overlooked, additional source of method variance, which leads to an increase in the confidence range for country mean estimates. In this respect, the findings from this paper are intended to provide an initial impetus not to overrate the significance of very small differences in means or, more generally, small effects from statistical analyses of LSA studies.

Abbreviations

FIML	Full information maximum likelihood
IRT	Item response theory
LI	Limited information
LSA	Large-scale assessment
MCMLM	Mixed coefficients multinomial logit model
ML	Maximum likelihood
MML	Marginal maximum likelihood
OLS	Ordinary least squares
PISA	Programme for international student assessment
PCM	Partial credit model
PV	Plausible values

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-022-00129-5.

Additional file 1. Overview of additional information, files, and results in the OSF repository APM2003-2012.

Acknowledgements
Not applicable.

Author contributions
Both authors contributed to all parts of the paper.

Funding
This research received no external funding.

Availability of data and materials
The original data used in this paper are available from the OECD download pages at https://www.oecd.org/pisa/data/. The Data as well as the SPPS and R syntaxes used in the present analysis are available from the Open Science Foundation (OSF) repository at https://osf.io/ubvaq/?view_only=99ba91912c0f450eaa198be51e88ded9.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Received: 18 March 2022 Accepted: 28 July 2022
Published online: 27 August 2022

References

Adams, R. J., Wilson, M., & Wc, Wang. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Anderson, C. A. (1961). Methodology of comparative education. *International Review of Education, 7*(1), 1–23. https://doi.org/10.1007/BF01416250

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2021) lme4: Linear mixed-effects models using 'Eigen' and S4. https://CRAN.R-project.org/package=lme4, R package version 1.1-27.1

Bolt, D. M. (2005). Limited- and full-information estimation of item response theory models. In R. P. McDonald, A. Maydeu-Olivares, & J. J. McArdle (Eds.), *Contemporary psychometrics: a Festschrift for Roderick P.* NJ: McDonald, Lawrence Erlbaum Associates.

Box, G. (1979) Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN (eds) Robustness in statistics: Proceedings of a workshop, Academic Press, New York, pp. 201–236

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics, 53*(2), 603. https://doi.org/10.2307/2533961

Choppin, B. H. (1968). Item bank using sample-free calibration. *Nature, 219*(5156), 870–872. https://doi.org/10.1038/2198700a0

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*(1), 5–32. https://doi.org/10.1007/BF02291477

Cohen, J., & Cohen, J. (Eds.). (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.
Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. Springer.
Edwards, M. C., & Orlando Edelen, M. (2009). Special topics in item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology*. SAGE Publications Inc.
Ertl, B., Hartmann, F. G., & Heine, J. H. (2020). Analyzing large-scale studies: Benefits and challenges. *Frontiers in Psychology, 11*(577), 410. https://doi.org/10.3389/fpsyg.2020.577410.
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.
Fischman, G. E., Topper, A. M., Silova, I., Goebel, J., & Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy, 34*(4), 470–499. https://doi.org/10.1080/02680939.2018.1460493
Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full Information methods. *Psychological Methods, 14*(3), 275–299. https://doi.org/10.1037/a0015825
Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), 305–322.
Glas, C. A. W., & Jehangir, K. (2013). Modeling country specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 97–115). CRC Press.
Grek, S. (2009). Governing by numbers: the PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23–37. https://doi.org/10.1080/02680930802412669
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
Heine, J.H. (2020). Untersuchungen zum Antwortverhalten und zu Modellen der Skalierung bei der Messung psychologischer Konstrukte. Monographie, Universität der Bundeswehr, München, Neubiberg, https://athene-forschung.unibw.de/132861
Heine, J.H. (2021). Pairwise: Rasch model parameters by pairwise algorithm. https://CRAN.R-project.org/package=pairwise, R package version 0.5.0-2
Heine, J. H., & Tarnai, C. (2015). Pairwise rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling, 57*(1), 3–36.
Henry, M. M. (1973). Methodology in comparative education: An annotated bibliography. *Comparative Education Review, 17*(2), 231–244.
Hopfenbeck, T., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing, 11*(2), 95–121. https://doi.org/10.1080/15305058.2010.529977
Husek, TR., & Sirotnik, K. (1967). Item sampling in educational research. CSEIP Occasional Report 2, University of California, Los Angeles, CA
Hutchison, D. (2008). On the conceptualisation of measurement error. *Oxford Review of Education, 34*(4), 443–460. https://doi.org/10.1080/03054980701695662
Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research, 58*(2), 139–148. https://doi.org/10.1080/00131881.2016.1165559
Johnston, J., & Dinardo, J. (1997). *Econometric methods* (4th ed.). McGraw Hill Book Company.
Kumar, A., & Dillon, W. R. (1987). The interaction of measurement and structure in simultaneous equation models with unobservable variables. *Journal of Marketing Research, 24*(1), 98–105. https://doi.org/10.2307/3151757
Lance, C. E., Cornwell, J. M., & Mulaik, S. A. (1988). Limited information parameter estimates for latent or mixed manifest and latent variable models. *Multivariate Behavioral Research, 23*(2), 171–187. https://doi.org/10.1207/s15327906mbr2302_3
Leamer, E., & Leonard, H. (1983). Reporting the fragility of regression estimates. *The Review of Economics and Statistics, 65*(2), 306–317. https://doi.org/10.2307/1924497
Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review, 75*(3), 308–313.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores. Addison-Wesley series in behavioral sciences: Quantitative methods*. Addison-Wesley Pub. Co.
MacCallum, R., Brown, M. W., & Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions, vol. 38*. Lawrence Erlbaum Associates.
Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center.
Martin, M. O., Von Davier, M., & Mullis, I. V. S. (2020). *Methods and Procedures: TIMSS 2019 technical report. Progress in international reading literacy study PIRLS*. TIMSS & PIRLS International Study Center.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272
Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika, 66*(2), 209–227.
Mazzeo, J., & von Davier, M. (2008). Review of the programme for international student assessment (PISA) test design: Recommendations for fostering stability in assessment results. Education Working Papers EDU/PISA/GB 28:23–24
Mazzeo, J., & Von Davier, M. (2013). Linking scales in international large-scale assessments. In L. Rutkowski, M. V. Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background technical issues and methods of data analysis* (pp. 229–258). CRC Press.
Mcaleer, M., Pagan, A. R., & Volker, P. A. (1985). What will take the con out of econometrics? *The American Economic Review, 75*(3), 293–307.

McArthur, D. L., & Wright, B. D. (1985). Bruce Choppin on measurement and education. *Evaluation in Education, 9*(1), 1–107. https://doi.org/10.1016/0191-765X(83)90005-8
McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates.
Michaelides, M. (2010). A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Frontiers in Psychology, 1*, 167. https://doi.org/10.3389/fpsyg.2010.00167
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196. https://doi.org/10.1007/BF02294457
Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x
Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika, 16*(1), 3–9. https://doi.org/10.1007/BF02313422
Mosteller, F. (1951). Remarks on the method of paired comparisons: II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika, 16*(2), 203–206.
Mosteller, F. (1951). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika, 16*(2), 207–218. https://doi.org/10.1007/BF02289116
OECD. (2005). *PISA 2003 technical report*. Organisation for Economic Co-operation and Development.
OECD. (2012). *PISA 2009 technical report*. PISA: OECD Publishing.
OECD (ed) (2014). PISA 2012 results: What students know and can do (Vol. I), PISA, vol I, revised edition. OECD Publishing, Paris
OECD. (2014). *PISA 2012 technical report*. PISA: OECD Publishing.
OECD. (2014). *PISA 2012 technical report*. PISA: OECD Publishing.
OECD. (2017). *PISA 2015 technical Report*. PISA: OECD Publishing.
OECD (2020). How to prepare and analyse the PISA database. https://www.oecd.org/pisa/data/httpoecdorgpisadatabase-instructions.htm
OECD (2021a). PISA 2018 Technical Report - PISA. https://www.oecd.org/pisa/data/pisa2018technicalreport/
OECD (2021b) Programme for international student assessment—data. https://www.oecd.org/pisa/data/
OECD, Adams R.J. (2009). *PISA 2006 technical report*. OECD Publishing.
OECD, Adams, R., Wu, M., (Ed.). (2002). *PISA 2000 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), 1–21. https://doi.org/10.1080/15305058.2013.825265
Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423–452. https://doi.org/10.1177/0013164413504926
Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340. https://doi.org/10.1126/science.abd3300
R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. No. 1 in Studies im mathematical psychology, Danmarks pdagogiske Institut, Kopenhagen
Robitzsch, A. (2020). About still nonignorable consequences of (partially) ignoring missing item responses in large-scale assessment. OSF Preprints 20 October 2020, https://doi.org/10.31219/osf.io/hmy45
Robitzsch, A. (2021). A comprehensive simulation study of estimation methods for the Rasch model. *Stats, 4*(4), 814–836. https://doi.org/10.3390/stats4040048
Robitzsch, A. (2021). 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. (2021). 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. (2022). Exploring the multiverse of analytical decisions in scaling educational large-scale assessment data: A specification curve analysis for PISA 2018 mathematics data. *European Journal of Investigation in Health, Psychology and Education, 12*(7), 731–753.
Robitzsch, A. (2022). On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy, 24*(6), 760. https://doi.org/10.3390/e24060760
Robitzsch, A., & Lüdtke, O. (2021). Reflections on analytical choices in the scaling model for test scores in international large-scale assessment studies. PsyArXiv 31 August 2021, https://doi.org/10.31234/osf.io/pkjth
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., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen [Relevance of item selection and model selection for assessing the development of competencies: The development in reading competence in primary school students]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43*(4), 213–227. https://doi.org/10.1026/0049-8637/a000052
Robitzsch, A., Kiefer, T., & Wu, M. (2021) TAM: Test analysis modules. https://CRAN.R-project.org/package=TAM, R package version 3.6-45
Rose, N., & von Davier, M. (2010). Xu X (2010) Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series, 1*, i–53. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
Rutkowski, D., & Rutkowski, L. (2021). Running the wrong race? The case of PISA for development. *Comparative Education Review, 65*(1), 147–165. https://doi.org/10.1086/712409
Rutkowski, D., Rutkowski, L., & Liaw, Y. L. (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? *Educational Measurement: Issues and Practice, 37*(4), 40–48. https://doi.org/10.1111/emip.12225
Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement, 48*(3), 293–312. https://doi.org/10.1111/j.1745-3984.2011.00144.x
Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education, 27*(2), 115–132. https://doi.org/10.1080/08957347.2014.880440
Rutkowski, L., Rutkowski, D., & Zhou, Y. (2016). Item calibration samples and the stability of achievement estimates and system rankings: Another look at the PISA model. *International Journal of Testing, 16*(1), 1–20. https://doi.org/10.1080/15305058.2015.1036163
Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education: Principles, Policy & Practice, 26*(6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.
Singer, J. D., & Braun, H. I. (2018). Testing international education assessments. *Science, 360*(6384), 38–40. https://doi.org/10.1126/science.aar4952
Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Springer, Wien.
UNESCO, (Ed.). (2019). *The promise of large-scale learning assessments: acknowledging limits to unlock opportunities*. Paris: UNESCO Institute for Education.
van den Heuvel-Panhuizen, M., Robitzsch, A., Treffers, A., & Köller, O. (2009). Large-scale assessment of change in student achievement: Dutch primary school students' results on written division in 1997 and 2004 as an example. *Psychometrika, 74*(2), 351. https://doi.org/10.1007/s11336-009-9110-7
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
Wu, M.L., Adams, R.J., Wilson, M., Haldane, S.A. (2012). ACER ConQuest: Generalised item response modeling software. Version 3.0
Fitting the structured general diagnostic model to NAEP data (ETS RR-08-27). *ETS Research Report Series, 1*, i–18. https://doi.org/10.1002/j.2333-8504.2008.tb02113.x

Publisher's Note