Psychological Test and Assessment Modeling, Volume 60, 2018 (2), 241-263

Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data

Matthias Trendtel[1,2] *& Alexander Robitzsch*[3,4]

Abstract

Previous studies have repeatedly demonstrated the existence of item position effects in large-scale assessments. Usually, items are answered correctly more often when administered at the beginning of a test compared to at the end of a test. In this article, the aspects of item position effects that are investigated are their pattern, whether they remain stable over time, and whether they are affected by changes in the test administration mode. For this purpose, a Bayesian item response model for modeling item position effects is proposed. This model allows for nonlinear position effects on the item side and linear individual differences on the person side. A full Bayesian estimation procedure is proposed as well as its extension to data collected from stratified clustered samples. The model was applied to the reading data collected in the 2009, 2012, and 2015 cycles of the Programme for International Student Assessment (PISA) for six countries.

Keywords: Item position effects, Bayesian IRT model, MCMC estimation, stratified clustered sample, PISA

[1] *Correspondence concerning this article should be addressed to:* Matthias Trendtel, Center for Research on Education and School Development, Vogelpothsweg 78, 44227 Dortmund, Germany; email: matthias.trendtel@tu-dortmund.de

[2] Main parts of this paper were completed while the first author was employed by the Federal Institute for Education Research, Innovation and Development of the Austrian School System, Salzburg, Austria

[3] Leibniz Institute for Science and Mathematics Education, Kiel, Germany

[4] Centre for International Student Assessment, Germany

Introduction

Large-scale assessments (LSAs) in the field of education aim to monitor student competencies and to examine them as indicators of school-system quality from an output perspective. LSAs describe specific student abilities, how and why students at certain schools show certain levels of abilities, or even how and why the students of whole countries show different levels of abilities from those of other countries. National LSAs are typically aligned with national curricula and aim to provide information about what students of the respective country know and are able to do with regard to the respective curriculum. International LSAs assess students from different countries and measure the abilities or skills that are assumed to be crucial for students' participation in society. The Programme for International Student Assessment (PISA; OECD, 2017), for instance, "assesses the ability of 15-year-olds to apply their knowledge in reading, mathematics, and science to real-life problems, rather than the acquisition of specific curriculum content" (Kirsch, Lennon, von Davier, Gonzalez, & Yamamoto, 2013, p. 17). International LSAs make comparisons between the participating countries' educational outcomes possible.

Whether national or international, LSAs provide researchers and policy makers with benchmarks for judging educational outcomes (i.e., those of an education system) within a national or international frame of reference. In general, LSAs are conducted iteratively. PISA, for instance, takes place every three years. Providing information about educational outcomes and their development over a longer period of time makes it possible to monitor education systems. By interpreting the findings of LSAs, policy makers gain insights into the functionality, productivity, or even equity of a country's education system. Whenever these insights are relevant for decision making, the validity of those interpretations must be confirmed.

Item position effects in LSAs

In order to meet the abovementioned objectives of LSAs (i.e., valid comparisons between countries, reports of ability trends within countries over time), standardized achievement tests are applied to measure the ability construct under investigation. These tests, consisting of a set of items, are administered to students. The observed performance a student shows in solving these items indicates the level of his or her ability, which the test is supposed to measure. To derive measures of the ability of interest from the observed performance, several statistical models can be used. The most prevalent models are grounded in the item response theory (IRT; e.g., van der Linden & Hambleton, 2013). In all of these models, the assumption has to be made that all relevant factors that have an influence on a student's test performance are captured by the model.

At the item level this means that administered items are assumed to function invariantly regarding some of their characteristics, if these characteristics are not captured by the model. In this respect, 'functioning' can be understood as the influence of the item and its characteristics on the way in which it is responded to by students. The influence that an item characteristic has on the response behavior should correspond to at least some parts of the construct that the item is intended to measure. In this case, the item characteristic causes construct-relevant variance in the response behavior. However, there might also be some item characteristics that do not function invariantly and that cause variance in the response behavior that is not intended to be measured. This kind of variance is called construct-irrelevant variance. Construct-irrelevant variance can affect the derivation of test scores and thereby threaten the validity of the interpretations of these test scores (Messick, 1995).

Effects that represent construct-irrelevant variance can be traced back to changes in the test administration or in test contexts, for example. Probably the most well-known example of a validity threat to the interpretation of test scores is the 1986 NAEP reading anomaly described by Zwick (1991). A suspiciously large decline in the estimated reading proficiency from 1984 to 1986 could be traced back to changes in test contexts. Furthermore, a well-known item characteristic that influences response behavior is the position within a sequence of items in which an item is administered to a student (e.g., Mollenkopf, 1950). Since variance caused by so-called item position effects (IPEs) is usually not intended to be measured, these effects might threaten the validity of test score interpretations.

Typically, in LSAs, several test forms that each consist of different items are administered and each item occurs in different positions within the different test forms. Consequently, IPEs should always be taken into consideration in LSAs. Furthermore, LSAs are low-stakes tests, which means that the result of the test does not have any personal consequences for the students who have taken it. The fact that test results have no individual impact might reduce the students' test-taking motivation and, consequently, the effort that the students invest in the test. Wise and DeMars (2005) examined the relationship between test-taking motivation and test scores. They concluded that when students' abilities are assessed by a low-stakes test, there are indications that these abilities are substantially underestimated due to students' low test-taking motivation, which leads to low test-taking effort. Weirich, Hecht, Penk, Roppelt, and Böhme (2016) investigated the degree to which IPEs are moderated by test-taking effort and/or changes in test-taking effort. Even though they could not completely explain the variances found in IPEs with the variances found in test-taking effort, their findings underline the fact that LSAs, due to their low-stakes character and the resulting lower test-taking effort, are especially susceptible to IPEs. The presence of IPEs in LSAs has already been shown in several studies. We refer the reader to the review, for instance, in Bulut, Quo, and Gierl (2017).

When analyzing data derived from LSAs in which IPEs may be present, two important points/steps are (1) the equating of test scores derived from different test forms and (2) the linking of test scores derived from different assessments. First, in order to map the observed scores obtained in different test forms on a common scale, it is often assumed that the respective item parameters function invariantly across test forms and, therefore, invariantly across positions. Second, in order to link the abilities derived from two different assessments, it is assumed that some sort of measurement invariance holds (e.g., Meredith, 1993). Some researchers postulated at least partial invariance (e.g., Byrne, Shavelson, & Muthén, 1989), others claimed that the items that are common to both assessments function invariantly on average in both assessments (e.g., Kolen & Brennan, 2004).

One option to deal with IPEs is to use an appropriate test design. Such test designs aim (amongst other things) to balance position effects across items. That is, every item is, on average, affected by the same IPEs. As a result, the derived item parameters that represent an item characteristic such as the average item difficulty (across item positions) show no or only minimal bias when estimated in the population. Item parameters are then interpreted as parameters that were derived from a test of a certain length (including other characteristics such as test breaks), in which the items occur in (approximately) every position with equal frequency (e.g., Frey & Bernhardt, 2012; Frey, Hartig, & Rupp, 2009).

LSAs generally aim to provide information about subpopulations, such as individual countries. Beyond the assumption that IPEs are averaged out in a population, one could question whether IPEs are invariant between subpopulations (e.g., across students, groups, countries, or over time; Meyers, Miller, & Way, 2009). In this case of invariance, a comparison of abilities between subpopulations is unaffected by IPEs even if they are not explicitly accounted for. However, it has been shown in the literature that this assumption is questionable and that IPEs seem to vary across persons (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Robitzsch, 2009), between countries (Debeer, Buchholz, Hartig, & Janssen, 2014; Le, 2009), and over time (Nagy, Haag, Lüdtke, & Köller, 2017). Therefore, the difference in the test scores of different subpopulations such as countries might partly depend on the position of the administered items in the test (Frey & Bernhardt, 2012; Nagy, Lüdtke, & Köller, 2016; Nagy et al., 2017).

Whereas IPEs and even changes in IPEs have been analyzed and reported in the literature quite extensively, there is only little knowledge about how stable these effects are for certain subpopulations over a longer period of time when assessments take place iteratively. In this article, we focus on the IPEs that occurred in the reading domain of the PISA assessments in 2009, 2012, and 2015, and we analyze the stability of these IPEs for six countries across time. We chose this domain because reading is regarded as a key competence that students require for their academic success (e.g., Connor et al., 2011). In conducting these analyses, we were particularly interested in the most appropriate specification of a statistical model for quantifying IPEs. It is frequently assumed that IPEs follow a linear curve when considered as a function of test duration. We questioned the specification of IPEs as linear effects and allowed them to function more flexibly.

Nonlinearity of item position effects

Previous research on IPEs in LSAs has mainly been based on the assumption that IPEs on the person and item side are linear (e.g., Nagy et al., 2017, 2016; Weirich et al., 2016; Debeer et al., 2014; Debeer & Janssen, 2013; Hartig & Buchholz, 2012). This assumption is mainly based on the observation of growing exhaustion during testing situations on the person side, which leads to a linear drop in performance. However, Debeer and Janssen (2013) discussed alternative forms of modeling IPEs, such as a quadratic function. Arguments supporting alternative modeling approaches stem from psychological research, where it has been shown that the drop in performance during testing situations does not perfectly follow a linear function: While the drop in performance was significantly lower at the beginning of a test, it accelerated at the end of a test. For instance, Schweizer, Schreiner, and Gold (2009) focusing on intelligence, found that a nonlinear "acceleration" function outperformed a linear one. This was also supported by the work of Kang (2014), who used the PISA 2009 data and examined how different representations of item positions (categorical, linear, and quadratic) affected the relationship between item difficulty and item position. For the domain of mathematics in vocational education, Frey, Bernhardt, and Born (2017) showed that IPEs followed a curvilinear function. Taking these findings together, it seems reasonable to assume that IPEs in PISA reading assessments are nonlinear. In more detail, the present paper investigates whether violations to the linearity of IPEs can be found in the PISA reading assessments. In addition to the pattern of IPEs within one assessment, we investigated whether IPEs varied within countries and across assessments.

Stability of item position effects across assessment cycles

PISA – as an example of a school-system monitoring study – is conducted in a regular three-year cycle in a multitude of countries. PISA assesses the domains of reading, mathematics, and science in each assessment. Therefore, it is ideally suited to study the stability of IPEs across countries and points of measurement. IPEs generally depend largely on the test context and the testing situation. For instance, IPEs tend to increase with test duration (e.g., OECD, 2017). Therefore, IPEs can be expected to be greater in magnitude for longer tests than for shorter ones. For two similar tests of equal length, one would expect IPEs to be similar in magnitude. The testing time in PISA is 120 minutes in all cycles. In addition, other aspects of the test context, such as different combinations of the domains to be measured (Zwick, 1991) or the item format, can impact on IPEs. Le (2009), for instance, showed for the PISA 2006 science data that constructed-response items tended to exhibit larger IPEs than closed-response items (such as multiple-choice items).

IPEs and test context effects (e.g., Brennan, 1992) can often not be separated; IPEs are considered to be a specific type of test context effects (e.g., Nagy et al., 2016). However, if the test context and testing situation do not vary notably across assessments, it can also be assumed that IPEs do not vary. As the participating countries in PISA tests must administer the test in a standardized manner, it can be assumed that the IPEs across assessments are quite stable.

A relatively new development within international LSAs is the change in the test administration mode. Recently conducted LSAs were designed as computer-based assessments (CBAs). In PISA 2015, CBA replaced paper-pencil testing (PPT) for the majority of countries (OECD, 2017). This is a notable change in the testing situation. CBA makes it possible to administer the test fully adaptively or adaptively in multiple stages and provides greater flexibility in positioning items. The latter might be at least partly the reason for why IPEs tended to decline on a global level from PISA 2012 to the PISA field test in 2014 (OECD, 2015) and to the main assessment in 2015 (OECD, 2017). Taking these results into account, it can be assumed that, on the country level, IPEs in PISA 2015 (CBA) were smaller than in PISA 2012 (PPT).

Research questions

The presence of IPEs within educational and psychological assessments has been demonstrated in various contexts. Due to their low-stakes character, the results of LSAs might be especially affected by IPEs. IPEs might be a threat to the validity of comparisons between the school systems of different countries, and might thereby be a threat to the aims of LSAs. This study adds to the literature on IPEs by studying (1) the nonlinear patterns of IPEs, (2) their stability across countries, and (3) whether the internationally reported decline in IPEs at a global level can be replicated on the country level for PISA 2012 and 2015.

Summing up, the following research questions were investigated:

1. Did IPEs for the reading domain in PISA 2009, 2012, and 2015 follow a nonlinear curve within countries and across assessments?

2. Did IPEs for the reading domain in PISA 2009 and 2012 vary within countries and across assessments?

3. Did IPEs decline within countries from PISA 2012 to PISA 2015?

Taking theoretical assumptions and previous empirical work into account, the following hypotheses can be formulated. We assumed that IPEs within countries and across assessments would be nonlinear (Hypothesis 1), due to the nonlinear decrease in student performance, which has already been demonstrated. Moreover, we expected IPEs to be stable within countries and across the assessments 2009 and 2012 (Hypothesis 2), because PISA is administered according to a standardized assessment procedure. Considering IPEs for the reading domain in PISA 2012 and 2015, we expected to find a decline in IPEs within countries (Hypothesis 3) and, thereby, a replication of this decline on the global level.

Method

Modeling item position effects in the IRT framework

Item position is defined as the position in which an item is administered within a test form also referred to as booklet. In most test designs, booklets consist of different but fixed clusters that contain several items. Clusters can occur in different positions within a booklet. The item position is thus operationalized by the position of the cluster that contains the item.

Let $p = 1, ..., N; i = 1, ..., I$ and $t = 1, ..., T$ denote the indices representing persons, items, and item positions, respectively. Further, let Y_{pit} denote the binary random variable representing the response of person p to item i when administered in position t. Note that this notation is somewhat redundant since an item is usually administered to person p in one position only. However, to facilitate interpretation, the index t is kept. The first model (Model M0) used to assess individual position effects in Debeer et al. (2014) can be formulated as follows:

$$\text{Model M0:} \quad \mathit{probit}\,[P(Y_{pit} = 1)] = \theta_p + (t-1)\eta_p + b_i + (t-1)\beta \tag{1}$$

where b_i is a fixed effect representing the item-specific difficulty. θ_p is the random effect representing the ability. β is the average change in performance from one item position to the next, while η_p is the individual deviation from this average and is also called persistence (e.g., Hartig & Buchholz, 2012). A positive (negative) value of η_p increases (decreases) the probability of a correct response at position $t > 1$. Note that we used the $probit$ function instead of the $logit$ function as the link function for computational reasons, where $probit(x) = \Phi^{-1}(x)$ and Φ is the standard normal distribution function. The vector $(\theta_p, \eta_p)^T$ follows a bivariate normal distribution with mean vector $\boldsymbol{\mu}_{\theta,\eta} = \mathbf{0} = (0,0)^T$ and covariance matrix

$$\boldsymbol{\Sigma}_{\theta,\eta} = \begin{pmatrix} \sigma_\theta^2 & \\ \sigma_{\theta\eta} & \sigma_\eta^2 \end{pmatrix} \tag{2}$$

where σ_θ, σ_η, and $\sigma_{\theta\eta}$ are free parameters representing the standard deviation of θ and η, and the covariance between θ and η, respectively. Dependencies between the ability and the persistence are captured by $\sigma_{\theta\eta}$. Model M0 falls into the class of generalized linear mixed effects models (Albano, 2013; Debeer & Janssen, 2013; De Boeck & Wilson, 2004; Hartig & Buchholz, 2012; Weirich, Hecht, & Böhme, 2014). This model can also be specified as a multidimensional Rasch model, which allows linear constraints to be placed on item parameters (Adams, Wilson, & Wang, 1997).

We extended the Model M0 by introducing dimension-specific loadings a_i and g_i as well as position-specific item effects β_t allowing for nonlinear position-specific change on the item side (Debeer & Janssen, 2013; Hecht, Weirich, Siegle, & Frey, 2015). That is,

$$\text{Model M1:} \quad \mathit{probit}\,[P(Y_{pit} = 1)] = a_i\theta_p + g_i(t-1)\eta_p + b_i + \beta_t \tag{3}$$

where a_i is the item discrimination regarding the ability and g_i can be interpreted as the extent to which the performance in solving item i is affected by the persistence η. The parameter β_t represents the position-specific effect of position t on items and is allowed to deviate from linearity (see Le, 2009; Robitzsch, 2009). To ensure the identifiability of the model, β_1 is set to zero, σ_θ and σ_η are set to one and g_i are restricted to $g_i > 0$ for all $i = 1, ..., I$.

Model M1 can be considered to be a generalized nonlinear mixed effects model (Debeer & Janssen, 2013) or a two-dimensional item response model with constraints on item intercepts and item slopes. This model can be estimated with (marginal) maximum likelihood methods. However, we propose a Markov chain Monte Carlo approach, which makes the extension to more complicated models easier.

Markov chain Monte Carlo estimation

To fit the proposed model, a fully Bayesian approach was applied. A Markov chain Monte Carlo (MCMC) procedure using the Gibbs sampler (Gelfand & Smith, 1990) was implemented in R (R Core Team, 2016) by the authors.

The MCMC approach has found widespread use in the estimation of item response models (König & van de Schoot, 2017; Rupp, Dey, & Zumbo, 2004). Compared to maximum likelihood estimation methods, MCMC methods are relatively easy to implement for complex item response models (see Fox, 2010; Levy & Mislevy, 2016; Patz & Junker, 1999). The Gibbs sampling approach reduces complicated high-dimensional estimation problems to stepwise evaluations of lower dimensional posterior distributions (which are often chosen to be unidimensional). However, in contrast to maximum likelihood estimation, the Bayesian approach additionally requires the specification of prior distributions for all parameters (Hoff, 2009). After implementing the Gibbs sampler, we observed some instabilities in the estimation of the g_i parameters. To improve the stability, we decided to specify informative prior distributions for the g_i parameters. For all other parameters, noninformative (improper) prior distributions were chosen.

Our Gibbs sampler employs augmented latent data **z** which can be interpreted as continuous versions of corresponding dichotomous item responses **y** (Albert & Chib, 1993). If an item response is $y = 1$, it holds for the latent data that $z > 0$. In this data augmentation step, every dichotomous item response y_{pit} is mapped onto a real valued random variable z_{pit}. The use of augmented data considerably simplifies the Gibbs sampling steps (Levy & Mislevy, 2016). The joint posterior density of the parameters given the data is specified as

$$
\begin{aligned}
&p(\mathbf{a}, \mathbf{b}, \mathbf{g}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{z}, \boldsymbol{\Sigma}_{\theta,\eta}|\mathbf{y}) \\
&\propto \left[\prod_p \prod_{i(t)} P(Y_{pit} = y_{pit}|Z_{pit}) P(Z_{pit}|\theta_p, \eta_p, a_i, g_i, b_i, \beta_t)\right] \\
&\quad \times \left[\prod_p P(\theta_p, \eta_p|\boldsymbol{\Sigma}_{\theta,\eta})\right] \times \pi(\mathbf{a}, \mathbf{b}) \pi(\mathbf{g}) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\Sigma}_{\theta,\eta})
\end{aligned}
\tag{4}
$$

where the proportional sign is used since a normalizing constant is left out of the equation.

The MCMC sampling from the posterior distribution was divided into several steps for the different parameter groups of the model. Each of the sampling steps from the corresponding posterior distribution can be derived by using common Bayesian techniques (Hoff, 2009).

To illustrate how the sampling procedure works for single parameters, we describe the sampling step for the parameters (a_i, b_i). More details can be found in Glas (2012), for example. The parameters (a_i, b_i) are modeled as fixed effects with no prior distribution, that is, an improper prior distribution which is constant across the whole parameter range. In this step, for all other parameters $\theta_p, \eta_p, g_i, \beta_t$ as well as for the samples from the augmented data step z_{pit} current samples are available. These can be derived from previous sampling steps or, in the case of the first sampling cycle, can be the starting values. Let $p_1, ..., p_K$ denote the persons to whom item i was administered. Defining $Z^*_{pit} = Z_{pit} - g_i\eta_p - \beta_t = a_i\theta_p + b_i + \epsilon_{pi}$, the components of parameter vector $(a_i, b_i)^T$ can be considered to be coefficients of the normal regression of $\mathbf{Z}^*_i = (Z^*_{p_1 it}, ..., Z^*_{p_K it})^T$ on $\mathbf{X}_i = (\boldsymbol{\theta}_i, \mathbf{1})$, where $\boldsymbol{\theta}_i = (\theta_{p_1}, ..., \theta_{p_K})^T$ and $\mathbf{1}$ is the K-dimensional vector with components equal to one. The conditional distribution of (a_i, b_i) is given as a bivariate normal distribution:

$$
(a_i, b_i) \sim N\left((\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{Z}^*_i, (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\right) \tag{5}
$$

Drawing one sample from the conditional distribution described by (5) for each item i, forms the sampling step for the parameters (a_i, b_i).

It is known that the point estimator that results from MCMC estimation asymptotically converges to the maximum likelihood estimate (Gelman et al., 2014). To reveal this, we rewrote the joint posterior in (4). Let $\boldsymbol{\delta}$ denote the vector of all model parameters. Then, the joint posterior is given as

$$
p(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{z}, \boldsymbol{\delta}|\mathbf{y}) \propto \left[\prod_p P(\boldsymbol{y}_p|\boldsymbol{z}_p) P(\boldsymbol{z}_p|\theta_p, \eta_p, \boldsymbol{\delta}) P(\theta_p, \eta_p|\boldsymbol{\Sigma}_{\theta,\eta})\right] \times \pi(\boldsymbol{\delta}). \tag{6}
$$

It is evident that the posterior contains the likelihood part involving data from all persons and a joint prior distribution $\pi(\boldsymbol{\delta})$. If the latent data are integrated out, the posterior for model parameters $\boldsymbol{\delta}$ further simplifies to

$$p(\boldsymbol{\delta}|\mathbf{y}) \propto \left[\prod_p \int P(\boldsymbol{y}_p|\boldsymbol{z}_p)P(\boldsymbol{z}_p|\theta_p,\eta_p,\boldsymbol{\delta})P(\theta_p,\eta_p|\boldsymbol{\Sigma}_{\theta,\eta})d\mathbf{z}_p d\theta_p d\eta_p\right] \times \pi(\boldsymbol{\delta}) \quad (7)$$

If the prior is chosen to be improper (i.e., $\pi(\boldsymbol{\delta}) = 1$), the multivariate mode of the posterior distribution for $\boldsymbol{\delta}$ equals the maximum likelihood estimate (Gelman et al., 2014). Hence, MCMC estimation can be used to obtain estimates that are asymptotically equivalent to maximum likelihood estimates.

Extension to weighted clustered samples

In LSAs, samples are not drawn completely randomly, but rather in a hierarchical manner. Usually, in a first step, schools are drawn randomly, followed by a random selection of classes or students within schools (e.g., OECD, 2017). To account for this structure in the samples, an extension to weighted clustered samples is necessary as the statistical inference that assumes persons to be independent is biased and leads to underestimated standard errors (Rust, 2013). Therefore, the MCMC sampling steps must also involve sampling weights (e.g., Goldstein, 2011) and we incorporated the respective sample weight w_p for person p in the likelihood part of (7). This led to the posterior distribution described by

$$p(\boldsymbol{\delta}|\mathbf{y}) \propto \left[\prod_p \left(\int P(\boldsymbol{y}_p|\boldsymbol{z}_p)P(\boldsymbol{z}_p|\theta_p,\eta_p,\boldsymbol{\delta})P(\theta_p,\eta_p|\boldsymbol{\Sigma}_{\theta,\eta})d\mathbf{z}_p d\theta_p d\eta_p\right)^{w_p}\right] \times \pi(\boldsymbol{\delta}) \quad (8)$$

For the parameters (a_i, b_i), for instance, this means that samples are drawn from the conditional distribution

$$(a_i, b_i) \sim N\left((\mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{W}_i\mathbf{Z}_i^*, (\mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i)^{-1}\right) \quad (9)$$

where $\mathbf{W}_i$ is a diagonal matrix that has entries $w_{p_1}, ..., w_{p_K}$ on the diagonal.

MCMC sampling provides approximations for the posterior distributions of the respective parameters. Point estimates of parameters can be directly obtained as the mode or the mean of their corresponding marginal posterior distributions. Thus, the statistical inference, including sample weights, can be seen as a Bayesian variant of pseudo-likelihood estimation (Rabe-Hesketh & Skrondal, 2006).

In addition, the computation of standard errors should take the stratified clustered sampling into account (OECD, 2017). Therefore, standard errors are computed by applying resampling methods (balanced repeated replicate weights). In the case of PISA, the resampling design consists of $R = 80$ replication weights (OECD, 2017). Applying the

MCMC estimation R-times to the data with the respective replication weights results in R replicated statistics, from which a covariance matrix $\boldsymbol{V}_{\hat{\boldsymbol{\delta}}}$ for a vector of estimated parameters $\hat{\boldsymbol{\delta}}$ can be calculated (e.g., Kolenikov, 2010). MCMC estimation with replicated weights is expected to perform similarly to pseudo-likelihood estimation with sample weights.

The covariance matrix $\boldsymbol{V}_{\hat{\boldsymbol{\delta}}}$ can be used to test the linear hypotheses of the form $\boldsymbol{C\delta} = \boldsymbol{d}$. The respective test statistic $X^2 = (\boldsymbol{C}\hat{\boldsymbol{\delta}} - \boldsymbol{d})'\boldsymbol{V}_{\hat{\boldsymbol{\delta}}}^{-1}(\boldsymbol{C}\hat{\boldsymbol{\delta}} - \boldsymbol{d})$ is χ^2 distributed with $df = rank(\mathbf{C})$ degrees of freedom. When testing hypotheses regarding single parameters, the derived standard errors can be used to construct the usual Z tests.

Data

The PISA samples of Austria, Finland, Germany, Japan, Portugal, and Turkey were analysed. Regarding the internationally reported results for 2009 (OECD, 2010), the sample comprised two countries that were significantly above (Finland and Japan), two countries that were significantly below (Austria and Turkey), and two countries that were not significantly different from the OECD average (Germany and Portugal).

We used the published reading data of PISA 2009, 2012, and 2015 (OECD, 2012, 2014, 2017). We included only students to whom booklets containing reading items were administered and for whom at least one valid response to a reading item was observed. The resulting sample sizes for each assessment and each country are shown in Table 1.

Table 1:
Number of students included in the analyses of the PISA 2009, 2012, and 2015 reading data from Austria, Finland, Germany, Japan, Portugal, and Turkey.

Country	2009	2012	2015	Country	2009	2012	2015
Austria	6,475 (470)	3,274	2,417	Japan	6,065 (520)	4,338	2,244
Finland	5,806 (536)	6,058	2,007	Portugal	6,294 (489)	3,923	2,513
Germany	4,797 (497)	3,332	2,177	Turkey	4,982 (464)	3,350	2,006

Note. Values in parentheses represent the country specific mean performance on the overall reading scale reported in OECD (2010).

In all PISA cycles, booklets contain four clusters of items, resulting in 4 different operationalized item positions. One cluster consists of items of solely one domain. In 2009, reading was the major domain in PISA and, consequently, all 13 booklets contained at least one cluster that consisted of reading items. In total, 101 reading items[5] were administered. In 2012 and 2015 reading was a minor domain. In 2012, 13 booklets were administered and nine booklets (booklets 2, 3, 4, 6, 8, 9, 11, 12, and 13) contained

[5] Two items in Austria and one item in Finland, one in Germany, and one in Japan were excluded due to national item deletion (OECD, 2012)

reading clusters, with a total of 44 reading items[6]. In 2015, PISA was conducted as a CBA in most of the participating countries (and in all of the six countries that we are focusing on here). This made it possible to administer more booklets than in PPT. Therefore, in 2015, 66 booklets (ignoring different versions of booklets consisting of different combinations of science clusters) were administered and 36 booklets contained reading clusters, with a total of 88 reading items[7].

Some of the items had originally been scored polytomously with three categories (incorrect, partially correct, and full credit). These items were dichotomized, scoring the full credit as correct (1) and partially correct and incorrect answers as incorrect (0). This applied to eight reading items in 2009, one reading item in 2012, and seven reading items in 2015. All missing responses that were coded as omitted were treated as incorrect responses. All missing responses that were coded as not reached and all other missing responses were treated as missing by design.

Procedure

We analyzed the reading data from PISA 2009, 2012, and 2015 separately for each country and each cycle. To derive estimates for the IPEs, Model M1 was fitted to the data by applying the described MCMC sampling method. The sampling procedure had a burn-in period of 25,000 iterations and the whole chain was 50,000 iterations long. Reported estimates for all parameters are the respective expected a posteriori (EAP), that is, the expected values of the respective posterior distributions. To assess the convergence of the MCMC algorithm, the potential scale reduction factor (PSRF), proposed by Gelman and Rubin (1992), was calculated, as were the effective sample sizes to ensure a sufficiently small MCMC error (Hoff, 2009). In order to calculate the PSRF, we divided the long chain into 4 chains of equal length. The PSRF never exceeded 1.05 for any estimated parameter and the effective sample size was always greater than 150. Further, we inspected the traceplots of every estimated parameter and the respective autocorrelation function with an increasing lag between the successive draws. All models met the respective diagnostic criteria.

To compare the variance explained by θ and η, respectively, one would have to impose some normalization restrictions on the a_i and g_i parameters in Model M1. Then, the parameters σ_θ and σ_η could be free parameters and the model could remain identifiable. Another way would be to consider the respective mean of the parameters a_i and g_i. Of course, these two values are conceptually not the same as the respective standard deviations. However, the balanced booklet design used in PISA (e.g., Frey & Bernhardt, 2012) and the randomized assignment of students to booklets leads to all items making approximately equally weighted contributions to the respective factors. Therefore, the respective means of a_i and g_i can be considered as good approximations.

To specify the aforementioned prior distribution for the g_i parameters, we used the results

[6] One item was excluded in Austria, in Finland, and in Germany due to national item deletion (OECD, 2014)

[7] One item was excluded in Turkey because the relative frequency of correct responses was < 0.01.

provided by Debeer et al. (2014) for PISA 2009. The authors analyzed the PISA 2009 reading data from 65 countries, including the countries we are focusing on in the present article. For their analyses, they fitted – amongst others – the Model M0 to the data and presented estimates for the standard deviation of η denoted by $\hat{\sigma}_{\eta}^{prior}$ for each country. It should be noted that the estimates provided by Debeer et al. (2014) were calculated on the logit metric. Therefore, we transferred the estimates to the probit metric using the factor $1/1.7$ (e.g., Lord & Novick, 1968). For each country we specified the prior distribution of all parameters g_i as a normal distribution with a mean of $\hat{\sigma}_{\eta}^{prior}/1.7$ and a standard deviation of 0.2. This resulted in a weakly informative prior, which sufficiently stabilized estimation. For the analyses of the 2012 data (2015 data), we used the respective country-specific mean and the standard deviation of the estimated parameters that resulted from fitting Model M1 to the 2009 data (2012 data) for each country.

Standard errors were calculated using the replicate weights provided in the published datasets. To answer Research Question 1, respective χ^2 tests were conducted, with the null hypothesis that linearity holds for the parameters β_2, β_3 and β_4 (note that β_1 is fixed to zero). As mentioned before, this can be done be specifying a proper matrix $\boldsymbol{C}$. We choose the equations $\beta_3 = 2\beta_2$ and $\beta_4 = 3\beta_2$ representing linearity. This leads to $\boldsymbol{C} = \begin{pmatrix} 2 & -1 & 0 \\ 3 & 0 & -1 \end{pmatrix}$ and linearity holds if $\boldsymbol{C}(\beta_2, \beta_3, \beta_4)^T = 0$.

For the other research questions, the respective Z tests were conducted separately for each pair of parameters. To make the IPEs comparable across two different years of assessment, we used approximations of the standardized IPEs $\beta_t^* = \beta_t/M(a_i)$. For Research Question 2, the null hypothesis was $\beta_t^{*^{2009}} = \beta_t^{*^{2012}}$ for $t = 2, ..., 4$ where $\beta_t^{*^{2009}}$ and $\beta_t^{*^{2012}}$ denote the approximation of the standardized parameter β_t^* for 2009 and 2012, respectively, with a corresponding two-sided Z test. The null hypothesis for Research Question 3 was $\beta_t^{*^{2012}} \leq \beta_t^{*^{2015}}$ where $\beta_t^{*^{2015}}$ denotes the approximation of the standardized parameter β_t for 2015, with a corresponding one-sided Z test. We report significance starting from a nominal alpha level of .05.

Results

Table 2 summarizes the results of Model M1, which was fitted separately to the data of the six selected countries for PISA 2009, 2012, and 2015. The mean of the estimated item discriminations $M(a_i)$, the mean of the estimated item specific loadings $M(g_i)$ on η, the correlation $\rho_{\theta,\eta}$ between θ and η, and the mean of estimated item difficulty $M(b_i)$ are shown.

The means $M(a_i)$ and $M(g_i)$ can be seen as proxy variables for the standard deviation of θ and η, respectively (with the abovementioned caveats). Considering $M(g_i)/M(a_i)$, the values ranged from around one sixth (Finland in 2012) to around one forth (Japan in 2015). This means that, for the case of Finland in 2012, the variance explained by η

was as high as one sixth of the amount of variance explained by θ. For most countries, the estimated correlation $\rho_{\theta,\eta}$ was positive for the years 2009 and 2012 and, in some cases, even significantly so. Turkey and Portugal were exceptions for the year 2009 and 2012, respectively, where the estimated correlations were slightly negative. A positive correlation between θ and η means that students with a higher level of ability showed lower declines in performance over the course of the test compared to students with a lower level of ability. However, except for Austria, in 2015 all estimated correlations were negative, which means that students with a higher level of ability showed higher declines in performance over the course of the test compared to students with a lower level of ability. In 2015, another mode of test administration was introduced, that is, CBA replaced PPT; maybe this is one reason for the different correlations obtained when comparing PISA 2009 and 2012 with PISA 2015.

The IPEs, that is, the mean change in difficulty for each item position, are also presented in Table 2. Values displayed in boldface indicate a significant difference from zero, which is simultaneously the IPE of position one. All IPEs were consistently negative, which means that, compared to position one, the difficulty of an item increased when it was administered in position two, three, or four.

Regarding position two, except for Finland in 2009 and 2015, the effects of this item position were significantly different from zero. For item position three, there was a somewhat more heterogeneous pattern: For Austria in 2012, and for Finland, Germany, and Japan in 2015, no significant difference was observed between position one and position three. Furthermore, for Austria and Japan in 2012, as well as for Austria, Finland, Germany, Japan, and Turkey in 2015, there was even a decline in IPEs from position two to position three. It should be noted that there was a break in testing between position two and three. The different patterns in the IPEs found for position three might be due to the extent to which students from different countries recovered in the testing break. Possibly, the change from PPT to CBA also had an influence on how students recovered during the break. Considering item position four, all estimated mean changes in difficulty were significantly different from zero. Further, compared to the effects of position three, all effects were higher in magnitude for position four.

Again, using $M(a_i)$ as a proxy variable for the standard deviation of θ, we were able to approximate standardized IPEs. Regarding position two, standardized IPEs ranged from -0.031 SD for Finland in 2009 and 2012 to -0.281 SD for Japan in 2012. The estimated standardized IPEs for position three ranged from -0.002 SD for Finland in 2015 to -0.356 SD for Japan in 2009 and, for position four, the effects ranged from -0.100 for Germany in 2015 to -0.523 SD for Turkey in 2009. For Turkey in 2009, this means a decline in performance from position one to position four that could be expected between two individuals who are about one half of a SD apart regarding their ability.

Table 2:
Summary of the estimated parameters a_i, g_i, $\rho_{\theta,\eta}$ and b_i as well as the estimated item position effects for the PISA 2009, 2012, and 2015 reading data from Austria, Finland, Germany, Japan, Portugal, and Turkey.

	Year	$M(a_i)$	$M(g_i)$	$\rho_{\theta,\eta}$	$M(b_i)$	β_2	β_3	β_4
Austria	2009	0.750	0.134	**0.197** (0.076)	0.239	−**0.064** (0.018)	−**0.136** (0.019)	−**0.248** (0.018)
	2012	0.733	0.131	0.099 (0.107)	0.410	−**0.067** (0.033)	−0.053 (0.033)	−**0.171** (0.025)
	2015	0.699	0.135	0.152 (0.127)	0.433	−**0.159** (0.024)	−**0.108** (0.033)	−**0.203** (0.033)
Finland	2009	0.650	0.116	**0.235** (0.083)	0.669	−0.020 (0.017)	−**0.056** (0.016)	−**0.161** (0.018)
	2012	0.664	0.111	**0.252** (0.108)	0.586	−**0.065** (0.029)	−**0.098** (0.033)	−**0.143** (0.020)
	2015	0.770	0.133	−0.171 (0.133)	0.616	−0.024 (0.021)	−0.001 (0.036)	−**0.078** (0.038)
Germany	2009	0.707	0.133	**0.244** (0.063)	0.453	−**0.044** (0.018)	−**0.147** (0.018)	−**0.244** (0.018)
	2012	0.725	0.141	0.017 (0.126)	0.576	−**0.071** (0.033)	−**0.158** (0.035)	−**0.274** (0.035)
	2015	0.702	0.136	−0.152 (0.099)	0.523	−**0.079** (0.024)	−0.014 (0.030)	−**0.070** (0.033)
Japan	2009	0.699	0.138	**0.353** (0.051)	0.627	−**0.130** (0.014)	−**0.249** (0.014)	−**0.315** (0.015)
	2012	0.671	0.170	**0.265** (0.127)	0.782	−**0.194** (0.027)	−**0.167** (0.027)	−**0.271** (0.026)
	2015	0.646	0.160	−0.058 (0.078)	0.631	−**0.081** (0.024)	−0.051 (0.033)	−**0.158** (0.035)
Portugal	2009	0.680	0.113	**0.154** (0.062)	0.400	−**0.092** (0.014)	−**0.193** (0.013)	−**0.321** (0.016)
	2012	0.726	0.155	−0.021 (0.062)	0.449	−**0.149** (0.025)	−**0.215** (0.034)	−**0.373** (0.025)
	2015	0.706	0.157	−**0.230** (0.092)	0.473	−**0.134** (0.023)	−**0.084** (0.038)	−**0.153** (0.040)
Turkey	2009	0.629	0.120	−0.004 (0.082)	0.168	−**0.036** (0.016)	−**0.161** (0.016)	−**0.329** (0.021)
	2012	0.635	0.127	0.110 (0.078)	0.279	−**0.095** (0.022)	−**0.123** (0.027)	−**0.319** (0.023)
	2015	0.597	0.131	−0.064 (0.092)	−0.004	−**0.139** (0.024)	−**0.059** (0.026)	−**0.147** (0.028)

Note. $M(a_i)$ = mean of the estimated discrimination parameters a_i. $M(g_i)$ = mean of the estimated loading parameters g_i. $\rho_{\theta,\eta}$ represents the estimated correlation between θ (ability) and η (persistence). $M(b_i)$ = mean of the estimated item difficulty parameters b_i. β_2, β_3 and β_4 are the position specific average change in difficulty of position 2, 3 and 4, respectively. Values in parentheses represent the respective standard errors. Values displayed in boldface are significantly different from zero on a 5 % alpha level.

To answer Research Question 1, χ^2 tests were conducted, as described above. The respective test statistics X^2 and p-values are presented in Table 3. Regarding the α-level of .05 in 2009, there were significant deviations from linearity for the countries Finland, Japan, and Turkey. For 2012, this was true only for Japan and Turkey. In contrast, for 2015, significant nonlinearity was found for all countries except Finland. Altogether, Hypothesis 1 is only partly supported by the data.

Table 3:
Summary of the χ^2 tests regarding linearity in item position effects for the countries Austria, Finland, Germany, Japan, Portugal, and Turkey for the years 2009, 2012, and 2015.

	2009		2012		2015	
Country	X^2	p	X^2	p	X^2	p
Austria	2.881	.237	4.945	.084	**17.370**	.000
Finland	**13.617**	.001	0.474	.789	5.321	.070
Germany	5.343	.069	1.025	.599	**9.349**	.009
Japan	**10.223**	.006	**23.700**	.000	**10.814**	.004
Portugal	3.739	.154	3.617	.164	**9.831**	.007
Turkey	**31.615**	.000	**16.681**	.000	**20.532**	.000

Note. X^2 denotes the test statistic of the χ^2 tests with respective p-value denoted by p. Values displayed in boldface indicate statistical significance on a 5 % alpha level.

To investigate the extent to which IPEs were stable across different years of assessment, we conducted Z tests for the parameters of every position, estimated separately for PISA 2009 and 2012 for each country. The test statistics and p-values of the conducted tests are presented in the left part of Table 4.

Table 4:
Summary of the two-sided Z tests regarding stability between 2009 and 2012 (left part) and of the one-sided Z test regarding the decrease from 2012 to 2015 (right part) of the standardized item position effects β_2^*, β_3^* and β_4^* for the countries Austria, Finland, Germany, Japan, Portugal, and Turkey.

	2009 and 2012						2012 and 2015					
	β_2^*		β_3^*		β_4^*		β_2^*		β_3^*		β_4^*	
Country	T	p	T	p	T	p	T	p	T	p	T	p
Austria	0.100	.921	**−2.120**	.034	**−2.372**	.018	−2.405	.992	−1.260	.896	−0.996	.840
Finland	1.315	.188	1.100	.271	−0.743	.457	1.298	.097	**2.123**	.017	**1.961**	.025
Germany	0.702	.483	0.192	.848	0.598	.550	−0.235	.593	**3.047**	.001	**4.144**	.000
Japan	**2.311**	.021	**−2.381**	.017	−1.066	.286	**3.043**	.001	**2.592**	.005	**2.375**	.009
Portugal	1.755	.079	0.231	.817	0.993	.321	0.350	.363	**2.469**	.007	**4.490**	.000
Turkey	**2.202**	.028	−1.240	.215	−0.414	.679	−1.565	.941	1.541	.062	**4.272**	.000

Note. T denotes the test statistic of the Z tests with respective p-value denoted by p. Values displayed in boldface indicate statistical significance on a 5 % alpha level.

For Finland, Germany, and Portugal there is no significant indication for instability in IPEs. For Austria, there were significant differences in the IPEs regarding position three and four. For Turkey, the IPEs regarding position two and, for Japan those regarding position two and three seemed to vary significantly in magnitude when 2009 was compared to 2012. In summary, there were indications of instabilities in the IPEs for individual countries and specific positions. Therefore, Hypothesis 2 is only partly supported by the data.

To answer Research Question 3, we investigated whether there were significant decreases in IPEs from 2012 to 2015. Again, Z tests were conducted analogously as described in the previous section, that is, separately for every country and every position. The respective test statistics and p-values are presented in the right part of Table 4. Surprisingly, in Austria, the IPEs regarding all positions increased from 2012 to 2015 and, of course, there were no significant decreases. Regarding position four for all other countries, the respective IPE decreased significantly from 2012 to 2015. Regarding position three, this was true for all countries except Turkey. Japan was the only country for which the IPE regarding position two decreased significantly in magnitude. Hypothesis 3 is thus only partly supported by the data. Even though for the majority of comparisons there was a significant decrease in the IPEs from 2012 to 2015, this trend was observed more frequently in the later item positions.

Conclusion

We have introduced a flexible model for modeling nonlinear IPEs on the item side and linear IPEs on the person side. A fully Bayesian estimation procedure was proposed and its extension and application to data stemming from stratified clustered samples has been described and demonstrated. Applying the model to data from six countries from the PISA assessments in 2009, 2012, and 2015, we investigated (1) whether there were violations to the linear pattern of IPEs, (2) the extent to which IPEs were stable across two assessments when no substantial change in the assessment was obvious, and (3) whether the change in the mode of assessment, that is, from PPT to CBA, caused a decrease in the IPEs.

In ten of 18 cases, significant deviations from linearity were found. Therefore, we conclude that linearity cannot always be assumed and nonlinear patterns might have to be accounted for in IPEs in other contexts.

In 13 of 18 comparisons, we did not find significant change in the IPEs from 2009 to 2012. Therefore, it seems that IPEs function quite stably over time. However, we did find significant differences in IPEs in seven cases. Future research should investigate why these differences occurred in these cases in order to gain more insights into the stability of IPEs.

All of the countries that we examined, except for Austria, seemed to have benefited from the change to CBA, in that smaller IPEs were present, at least in the later parts of booklets. Four of the six countries showed significant decreases in IPEs regarding positions three and four. The magnitude in the decrease was more than one fourth of a SD for Turkey and position four. We were unable to find a reason for why this was not the case in Austria. However, further research should investigate the reasons for this contrasting pattern and aim to establish whether other countries show similar trends in IPEs.

Limitations and future developments Of course, our findings are limited to the six selected countries, the PISA assessment, the years 2009, 2012, and 2015, as well as to the domain of reading. In order to make our findings generalizable, we recommend further and deeper analyses of other countries, other assessments, other time points, and especially of other domains.

Our proposed IRT model for IPEs is limited to dichotomous item responses. However, an extension allowing for ordinal response data is possible. Further, we only investigated one literacy domain (i.e., reading) to analyze IPEs with a two-dimensional IRT model including the ability and the persistence factor. The model could be extended in order to simultaneously investigate three literacy domains (i.e., reading, mathematics, and science), which would result in three ability factors and three persistence factors.

The central assumption underlying the proposed model was a bivariate normal distribution for ability and persistence. However, the persistence component in the bivariate distribution could be assumed to follow a log-normal distribution multiplied by minus one, which only allows negative values. This model is related to a performance-decline item response model (Jin & Wang, 2014). Although the model is formulated to investigate declines in each item in a fixed-order test, the model could also be used for the case of IPEs in our study (concerning a non fixed-order test) and student performance declines could thus occur in the first, the second, third or the fourth position.

Implications Concerning the results for Austria and Finland, for instance, there is some disagreement between researchers on how to interpret the findings. First, some researchers argue that a comparison between these two countries is not valid because the country difference is not invariant across positions and measurement invariance is a requirement of the comparison of means (Kreiner & Christensen, 2014). However, IPEs are a form of differential item functioning at the country level that can be seen as part of the construct and should not be removed for cross-national comparisons (see Zwitser, Glaser, & Maris, 2017).

Second, some researchers argue that IPEs induce a violation to the assumed unidimensional scaling model because item parameters do not hold for items in all positions and additional local dependence is introduced due to the persistence factor. Simulation studies usually show that applying a model that ignores IPEs provides biased parameter estimates, while the model that includes IPEs leads to unbiased estimates (see, for example Debeer & Janssen, 2013). However, this reasoning is trivial as it simply shows that estimating a statistical model, which is used for generating data, provides acceptable estimates. In practice, the application of a statistical model to derive estimates of the ability distribution of a country requires the definition of factors that should be controlled for (and included in the model) or just nuisance factors (not to be included in the model, but to be taken note of as the reason for possible increases in standard errors). In the case of treating IPEs as a nuisance factor not included in the model, the violation of local dependence or the noninvariance of item parameters does not lead to biased parameter estimates because IPEs are considered to be part of the construct and should thus not be controlled for. Ability is seen as an index obtained as a summary of item responses in a unidimensional scaling model rather than as a student trait that causally influences manifest item responses (Borsboom & Wijsen, 2017). In this context, the well-defined statistical theory of misspecified models (the so-called quasi-maximum likelihood; see White, 1982) should be taken note of. Given the probability distribution of item responses in the population, a well-defined optimal parameter of the (misspecified) statistical model exists. Hence, the use of relatively simple IRT scaling models in large-scale assessments can be reasonably defended, but standard errors have to be computed with sandwich methods (see Ip, Molenberghs, Chen, Goegebeur, & De Boeck, 2013). Moreover, the full modeling of dependencies in large-scale assessment data does not seem feasible because of the existence of IPEs, differential item functioning (DIF) in subgroups such as gender, local dependence due to testlets, or multilevel DIF due to the nested structure of students within schools (see Robitzsch, 2009, for some applications).

Third, some researchers argue that using IRT models to analyze IPEs is helpful for the data analysis of data obtained by less than optimal test designs. For example, Nagy et al. (2017) described an application for investigating school-type differences in longitudinal performance over two waves of measurement. In their study, the test designs varied largely between the two waves, for example, the average position of the administered items differed between the two waves. Applying a model with IPEs makes the comparison of school types in both waves possible in a standardized condition, which is the consequence of including the control factor ”position” in the model (Nagy et al., 2017). The standardized condition in the application of Nagy et al. (2017) is the first position. Although the primary motivation behind this approach seems plausible to us, we think that it is not clear why a comparison between groups should be made in the first position and not in any other position in the test (or in an averaged position). Moreover, the adjustment of observed group differences by a model with IPEs relies on statistical assumptions. Maybe a better alternative could be to weight (or select or match) the data from the two waves in such a way that the test designs would be very similar and no statistical model with strong assumptions would have to be posed. Therefore, we believe that many problems concerning IPEs in large-scale assessments can be circumvented by using balanced test designs (Frey & Bernhardt, 2012) that represent IPEs in the ability construct in such a way that they provide unbiased parameter estimates.

Authors’ note

The authors are grateful to two anonymous reviewers for their time, comments and suggestions, as well as to Gráinne Newcombe for her valuable language editing.

References

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

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*(4), 408–426. doi: 10.1111/jedm.12026

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. doi: 10.2307/2290350

Borsboom, D., & Wijsen, L. D. (2017). Psychology's atomic bomb. *Assessment in Education: Principles, Policy & Practice*, *24*(3), 440–446. doi: 10.1080/0969594X.2017.1333084

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, *5*(3), 225–264. doi: 10.1207/s15324818ame0503_4

Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large-scale Assessments in Education*, *5*(8), 1–20. doi: 10.1186/s40536-017-0042-x

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. doi: 10.1037/0033-2909.105.3.456

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., … Schatschneider, C. (2011). Testing the impact of child characteristics x instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, *46*(3), 189–221. doi: 10.1598/RRQ.46.3.1

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523. doi: 10.3102/1076998614558485

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

De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 3–41). New York, NY: Springer.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer. doi: 10.1007/978-1-4419-0742-4

Frey, A., & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling*, *54*(4), 397–417.

Frey, A., Bernhardt, R., & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests [Handling of item positions effects in the development of computerized adaptive tests]. *Diagnostica*, *63*(3), 167-178. doi: 10.1026/0012-1924/a000173

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

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. doi: 10.2307/2289776

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). London, UK: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Glas, C. (2012). *Estimating and testing the extended testlet model* (LSAC Research Report Series; No. 12-03). Newtown, PA: LSAC.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester, UK: John Wiley & Sons. doi: 10.1002/9780470973394

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*(4), 418–431.

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*, *75*(6), 1021–1044. doi: 10.1177/0013164415573311

Hoff, P. D. (2009). *A first course in bayesian statistical methods*. New York, NY: Springer. doi: 10.1007/978-0-387-92407-6

Ip, E. H., Molenberghs, G., Chen, S.-H., Goegebeur, Y., & De Boeck, P. (2013). Functionally unidimensional item response models for multivariate binary data. *Multivariate Behavioral Research*, *48*(4), 534–562. doi: 10.1080/00273171.2013.796281

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

Kang, C. (2014). *Linear and nonlinear modeling of item position effects.* (Unpublished master's thesis). University of Nebraska.

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. Von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11). Dordrecht, Netherlands: Springer.

Kolen, M., & Brennan, R. (2004). *Test equating, linking, and scaling: Methods and practices*. New York: Springer. doi: 10.1007/978-1-4939-0317-7

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata Journal*, *10*(2), 165–199.

König, C., & van de Schoot, R. (2017). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, 1–24. doi: 10.1080/00131911.2017.1350636

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

Le, L. T. (2009). Effects on item positions on their difficulty and discrimination: A study in PISA science data across test language and countries. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 207–214). Tokyo, Japan: Universal.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: CRC Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi: 10.1007/BF02294825

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi: 10.1037/0003-066X.50.9.741

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education*, *22*(1), 38–60. doi: 10.1080/08957340802558342

Mollenkopf, W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. *Psychometrika*, *15*(3), 291–315. doi: 10.1007/BF02289044

Nagy, G., Haag, N., Lüdtke, O., & Köller, O. (2017). Längsschnittskalierung der Tests zur Überprüfung des Erreichens der Bildungsstandards der Sekundarstufe I im PISA-Längsschnitt 2012/2013 [Longitudinal IRT Scaling of Tests of the Educational Standards for Lower Secondary Level in the PISA Longitudinal Assessment 2012/2013]. *Zeitschrift für Erziehungswissenschaft*, *20*(2), 259–286. doi: 10.1007/s11618-017-0755-1

Nagy, G., Lüdtke, O., & Köller, O. (2016). Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment. *Psychological Test and Assessment Modeling*, *58*(4), 641–670.

OECD. (2010). *PISA 2009 Results: Executive Summary*. Paris, France: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. Paris, France: OECD Publishing.

OECD. (2014). *PISA 2012 Technical Report*. Paris, France: OECD Publishing.

OECD. (2015). *PISA 2015 field trial analysis report: Outcomes of the cognitive assessment (jt03371930)*. Paris, France: OECD Publications.

OECD. (2017). *PISA 2015 Technical Report*. Paris, France: OECD Publishing.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178. doi: 10.3102/10769986024002146

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805–827. doi: 10.1111/j.1467-985X.2006.00426.x

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodological challenges in the calibration of achievement tests]. In O. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss, & G. Walther (Eds.), *Bildungsstandards Deutsch und Mathematik* (pp. 42–106). Weinheim, Germany: Beltz.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling*, *11*(3), 424–451. doi: 10.1207/s15328007sem1103_7

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–153). Boca Raton, FL: CRC Press.

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly*, *51*(1), 47–64.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York, NY: Springer. doi: 10.1007/978-1-4757-2691-6

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, *38*(7), 535–548. doi: 10.1177/0146621614534955

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2016). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*(2), 115–129. doi: 10.1177/0146621616676791

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

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. doi: 10.1207/s15326977ea1001_1

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*(3), 10–16. doi: 10.1111/j.1745-3992.1991.tb00198.x

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