JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Summer 2016, Vol. 53, No. 2, pp. 152–171

A Comparison of Linking Methods for Estimating National Trends in International Comparative Large-Scale Assessments in the Presence of Cross-National DIF

Karoline A. Sachse, Alexander Roppelt, and Nicole Haag
Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Germany

Trend estimation in international comparative large-scale assessments relies on measurement invariance between countries. However, cross-national differential item functioning (DIF) has been repeatedly documented. We ran a simulation study using national item parameters, which required trends to be computed separately for each country, to compare trend estimation performances to two linking methods employing international item parameters across several conditions. The trend estimates based on the national item parameters were more accurate than the trend estimates based on the international item parameters when cross-national DIF was present. Moreover, the use of fixed common item parameter calibrations led to biased trend estimates. The detection and elimination of DIF can reduce this bias but is also likely to increase the total error.

Many large-scale testing programs use item response theory (IRT) and IRT-based equating methods to estimate trends over time. This procedure involves a comparison of latent variable means. For latent variable mean comparisons to be valid, the means must be located on the same scale. This can be ensured by using a linking procedure (Dorans, Pommerich, & Holland, 2007; Kolen & Brennan, 2004). Usually, the linking in educational large-scale assessments is established by employing a set of common items that are used on all test forms. Although the choice of the IRT model varies between different educational comparative large-scale assessments, most of these studies follow the same rationale and are therefore sensitive to the choice of linking method. Whereas prominent educational large-scale assessments such as the National Assessment of Educational Progress (NAEP; Allen, Donoghue, & Schoeps, 2001) or the Trends in International Mathematics and Science Study (TIMSS; Arora, Foy, Martin, & Mullis, 2009) use the 3PL IRT model, the Programme for International Student Assessment (PISA; OECD, 2012) employs a Rasch-type measurement model, namely, the multidimensional random coefficients multinomial logit model (Adams, Wilson, & Wang, 1997). As PISA's trend estimates are widely recognized and impact policy decisions in many countries (e.g., Cosgrove & Cartwright, 2014; Neumann, Fischer, & Kauertz, 2010), it is as important to evaluate the performance of PISA's linking methods as it is to evaluate the methods of other international large-scale assessments (ILSAs) that have a similar impact. Hence, the present study focuses primarily on trend estimation techniques employed by PISA and easy-to-implement variants.

Random or systematic inaccuracies can emerge from the many steps involved in the measurement and reporting of trends in student achievement. Three main sources of random error that may threaten the validity of the final results will be reviewed briefly. According to Wu (2010), the three main sources of error are (a) measurement error, (b) sampling error, and (c) equating error. Measurement error refers to random imprecision in the measurement process. Because repeated measurement reduces measurement error, error is usually reduced in psychometric testing by increasing test length. Sampling error occurs when the sample characteristics differ from the population characteristics. Aside from complicated sampling procedures in large-scale assessments, increasing the sample size is a standard procedure for diminishing random sampling error. It has been repeatedly shown that increasing the size of the random sample improves the estimation of linking coefficients (Hanson & Béguin, 2002; Kang & Petersen, 2012; Kim, Choi, Lee, & Um, 2008; Lei & Zhao, 2012; Tong & Kolen, 2007). The equating error component deserves increased attention because the given definitions are somewhat divergent. Whereas Kolen and Brennan (2004) base their definition of random equating error primarily on sampling error, the PISA technical reports, in line with Wu (2010), describe linking error or equating error as "the uncertainty in the transformation due to the sampling of the link items" (OECD, 2014b, p. 159). This corresponds to what Kolen and Brennan (2004) describe as one facet of systematic equating error: "In the common-item nonequivalent groups design, systematic error results if … the common items function differently from one administration to the next" (p. 24). If items always functioned in the same manner, it would make no difference which items were chosen as anchor items. In any case, an increase in equating error in the sense of Wu (2010) and the OECD (2014b) occurs when the assumption of measurement invariance is violated, which is called differential item functioning (DIF) in the IRT context. DIF occurs when the probability distributions of the response variable depend on latent trait and subgroup membership within the same population (Osterlind & Everson, 2009).

When an item functions differently in a specific country in comparison with the average item functioning across all other countries, the term cross-national DIF is applied. In recent years, cross-national DIF has become an increasingly recognized area of research in a variety of fields (e.g., Davidov, Schmidt, & Billiet, 2012; De Jong, Steenkamp, & Fox, 2007; Huggins, 2013), and many researchers have shown that it is reasonable to assume that not all common items exhibit measurement invariance across nations (Kreiner & Christensen, 2014; Oliveri & von Davier, 2011; Rutkowski & Svetina, 2014). Among the most likely reasons for cross-national DIF are (a) poor item translation, (b) cultural specificities (e.g., differences in connotative meaning), (c) differential curriculum coverage, and (d) differential content familiarity (Artelt & Baumert, 2004; Huang, 2010; van de Vijver & Tanzer, 2004). The consideration of only these few sources of cross-national DIF already makes it clear that cross-national DIF arises from many different sources that are quite difficult to disentangle (Grisay & Monseur, 2007) and that it is not easy to control.

Because a crucial point in trend estimation in cross-national achievement studies is that it actually comprises three linking steps rather than just one (OECD, 2012), it is suspected that national trend estimation in international large-scale assessments is particularly affected by cross-national DIF and its associated linking error. This

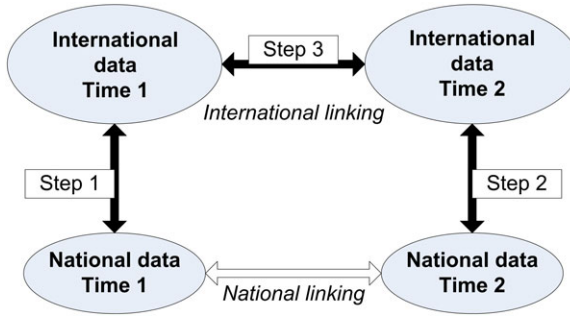

Figure 1. Steps involved in national and international linking.

increase in the number of steps originated from the goal of ranking countries on an international scale at each measurement occasion in addition to determining differences in student performance within one country across time. As depicted in Figure 1, first, international item parameters are estimated using the data from all countries, and national item parameters are linked to the international scale (step 1). Second, the procedure is repeated at time 2 (step 2). The international data at times 1 and 2 could, for example, correspond to the 2009 and 2012 PISA reading assessments, which comprise largely the same countries but not the same respondents within these countries. Third, international item parameters at time 2 have to be linked to the international scale at time 1 (step 3). The resulting transformation rule is then applied to the national parameter estimates. Every linking step can be a potential source of error.

Researchers have developed methods that allow for cross-national variation in item parameters (e.g., Fox & Verhagen, 2010) to prevent trend estimation from being affected by cross-national DIF. Moreover, Oliveri and von Davier (2011, 2014) proposed that all data be concurrently calibrated by using multiple group models and that misfitting international item parameters be replaced by item parameters that were based on national data for constructing the international scale, which resulted in an improved model-data fit. However, all these methods were not routinely applied to the aforementioned educational large-scale assessments until 2015 (in PISA, changes will be made from the 2015 cycle on). As another alternative that can be used to avoid the distorting influences of cross-national DIF, trend estimation based on national item parameters instead of international item parameters was established (Carstensen, 2013; Carstensen, Prenzel, & Baumert, 2008; Gebhardt & Adams, 2007). In *national linking* (see Figure 1), item parameters are estimated separately for each country and are aligned with the previously used test form in only one linking step (e.g., using concurrent calibration; Carstensen, 2013), whereas the international item parameters are not taken into account. Hence, the advantage of *national linking* is that a focal country's trend estimation is independent of cross-national DIF. However, this comes at the cost of international comparability.

To compute trends using international data in the PISA study (OECD, 2012), until 2015 national item parameters were linked to the international scale at all times

(steps 1 and 2 in Figure 1) via the fixed common item parameters (FCIPs) procedure implemented in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). FCIP means that item parameters are not estimated with the new data set but are fixed to the values obtained from the previous scaling. This is the most restrictive IRT linking method (von Davier & von Davier, 2007), a choice that might not be appropriate when cross-national DIF is assumed. Further, some research has indicated that the performance of FCIP is different from transformation methods such as mean-mean linking (MM; Loyd & Hoover, 1980), mean-sigma linking (Marco, 1977), or characteristic curve linking (Haebara, 1980; Stocking & Lord, 1983) and may yield biased results when there is a shift in the population that leads to changes in the latent ability distribution (Baldwin, Baldwin, & Nering, 2007; Hu, Rogers, & Vukmirovic, 2008; Keller & Keller, 2011; Kim, 2006; Paek & Young, 2005). This effect occurs most often when only a small number of common items are employed (Arai & Mayekawa, 2011) and it depends on the software implementation (Kang & Petersen, 2012; Kim, 2006; Paek & Young, 2005). According to Stocking and Lord (SL; 1983), an alternative to the FCIP procedure is characteristic curve linking. In the SL approach, data at both time points are scaled separately. Afterward, the item parameters obtained with one of the data sets are transformed to the scale of the other item parameters such that the squared difference between the test characteristic curves is minimized. Further, SL linking has been shown to be less sensitive to atypical test characteristics than other linking methods (e.g., Baker & Al-Karni, 1991) and to be more accurate than moment methods (Hanson & Béguin, 2002), which is in line with Kolen and Brennan's (2004) preference for characteristic curve methods. In all analyses, we also considered MM linking, which yielded results that were very similar to the results using SL linking. For reasons of clarity, we restricted our hypotheses and analyses to SL linking.

The present study focused on three alternative linking methods: (1) FCIP linking based on international item parameters because it was the operational practice in PISA in the past (until 2015), although it might not be the most appropriate approach in the presence of cross-national DIF because FCIP is the most restrictive IRT linking method; (2) SL linking based on international item parameters because SL is a frequently recommended IRT linking method (Kolen & Brennan, 2004), and we wanted to test whether it is an appropriate alternative to FCIP; and (3) SL linking based on national item parameters because it has sometimes been employed to avoid the negative effects of cross-national DIF on trend estimation.

First, we compared the trend estimates from linking based on national versus international data, that is, (3) versus (1) and (2). We expected national linking to be more accurate than international linking when cross-national DIF was present because linking based on international data relies on the assumption of measurement invariance between countries, but cross-national measurement invariance has repeatedly been challenged in the literature (e.g., Kankaraš & Moors, 2014; Kreiner & Christensen, 2014). By contrast, linking based on national data is independent of cross-national DIF. In cases without cross-national DIF, international linking was expected to be more accurate because of its larger sample size. Second, we compared a linking method using SL equating to one linking method using FCIP, that is, (2) versus (1). The international linking method employed by PISA until 2015 used

FCIP calibration, which has been shown to be biased under various conditions in previous studies, especially when a latent proficiency shift was present. We expected this bias to appear in our study as well, whereas SL linking was expected to remain unbiased.

Hence, we formulated two principal hypotheses:

1. In the presence of high cross-national DIF, SL linking that is based on national item parameters will lead to more accurate trend estimates than SL linking that is based on international item parameters.
2. The FCIP linking method that is based on international item parameters will lead to more biased trend estimates than the SL linking method that is based on international item parameters when a latent proficiency shift is present.

Method

In this study, we compared the trend estimates produced by three different linking methods using simulated data that were intended to mimic the designs of international or interfederal comparative large-scale assessments. For the purpose of trend estimation, we generated two international data sets representing time 1 and time 2, respectively. These consisted of 20 populations at each time point. At time 1, the countries' proficiency samples were drawn independently from a multivariate normal distribution with an expectancy of $\boldsymbol{\theta}_1$, a vector that consisted of equidistant values between -1 and 1, and an identity matrix as the covariance matrix $\boldsymbol{\Sigma}_\theta = I_{20}$. Depending on the simulated condition (see next section), the samples of the true proficiency values for time 2 were drawn either from the same distribution or from one with shifted true means. The items were arranged into eight blocks. The blocks contained 15 items each. Depending on the simulated condition, either four blocks (60 items) or two blocks (30 items) were common at both times 1 and 2 (i.e., 50% or 25% of the items were used as anchor items, respectively). Eight booklets were filled with two blocks each according to a balanced incomplete block design (Frey, Hartig, & Rupp, 2009). This design results in a personwise missing data ratio of 75%. Item difficulties were drawn blockwise from a multivariate normal distribution with the expectancy $\boldsymbol{\beta} = (-.7, -.5, -.3, -.1, .1, .3, .5, .7)^T$ and variance $\boldsymbol{\Sigma}_\beta = I_8$ in order to achieve good coverage across the proficiency range. A total of 1,000 data sets per condition were generated by changing the random number seed with the default random number generator in R.

Factors of Investigation

To test the hypotheses, five factors were fully crossed, resulting in $5 \times 2 \times 3 \times 2 \times 2 = 120$ conditions (Table 1). For every combination, three linking methods for computing the trend for one focal country were compared. The literature showed that FCIP is biased when the populations that have to be linked through a test differ in their true mean proficiencies (e.g., Keller & Keller, 2011). Thus, out of the 20 simulated countries, the 19th country was chosen as the focal country because its true proficiency mean was presumably always different from the international true proficiency mean.

Table 1
Summary of Manipulated Factors

Factor	No. of Levels	Level Description
Cross-national DIF	5	Country-specific item difficulty shift drawn from $N(0,.00)$, $N(0,.04)$, $N(0,.25)$, $N(0,.64)$, or $N(0,1.00)$; the same shift was applied to the anchor items at both times 1 and 2
DIF elimination	2	No items excluded versus items with $\|DIF\| > 1$ logit were excluded
Proficiency shift	3	No proficiency shift, national shift (only focal country), or international shift (all countries)
Sample size	2	$n = 500$ per country versus $n = 2{,}000$ per country
Anchor size	2	30 out of 120 total items versus 120 out of 240 total items

First, the cross-national DIF was manipulated. To find realistic DIF conditions, we examined cross-national DIF in the PISA 2009 and 2012 reading and mathematics assessment data (OECD, 2010, 2014a). To obtain the international item parameters and the cross-national DIF values, we followed the operational practice used by PISA until 2015 (OECD, 2012). First, we randomly sampled 500 persons per country in order to form the database for the international calibration. Then, we separately calibrated 63 single focal countries using their full sample sizes and compared the national anchor item parameters (which were centered around the scale mean of the national anchor item parameters) to the international anchor item parameters (which were centered around the scale mean of the international anchor item parameters), whereas only dichotomous items were considered. We found that the country-specific DIF variances ranged from .04 to .78 with an average DIF variance of .18 across all scales, years, and countries. Moreover, we found that it was realistic to assume very similar country-specific DIF at times 1 and 2. The correlations across all countries were distributed with a mean of .85 and a standard deviation of .10. For our experimental variation of simulation conditions, we chose five categories: (a) no DIF, (b) small DIF ($\sigma^2_{DIF} = .04$), (c) high DIF ($\sigma^2_{DIF} = .25$), (d) very high DIF ($\sigma^2_{DIF} = .64$), and (e) extreme DIF ($\sigma^2_{DIF} = 1.00$), according to and extending the Educational Testing Service's (ETS) classification for categorizing DIF effect variance (e.g., Penfield & Algina, 2006). The conditions with DIF variances of 1 and 0 were more extreme than the values identified in the PISA sample. However, we decided to include these DIF variances to show how the three linking procedures performed under extreme cross-national DIF conditions. The procedure was as follows. After the international item parameters were drawn, an additional item difficulty shift was added for each country and each item. This shift was drawn from a normal distribution with a mean of zero and a variance corresponding to the respective simulated

DIF condition (see Table 1). The same item difficulty shift was employed at both times for the anchor items. The difficulty shift for the nonanchor items varied between time 1 and time 2 because these items were not the same.

Second, in order to detect potential FCIP bias, the latent proficiency shift was varied across three levels: (a) no proficiency shift, that is, the proficiency expectancy remained the same as it had been at time 1; (b) national proficiency shift: only the focal country's mean true proficiency increased by .2 logits; (c) international proficiency shift: each country's mean true proficiency increased by a random value drawn from a normal distribution with a mean of zero and a standard deviation of .2 except for the focal country's mean true proficiency, which always increased by .2 logits.

Subsequently, three side factors were manipulated to ensure the validity of our results across different designs. These side factors consisted of DIF item elimination, the sample size, and the number and percentage of anchor items. In the DIF elimination conditions, all items exhibiting an absolute DIF greater than 1 logit were excluded from each link. This was done separately for all linking steps. The literature has provided sophisticated rules and strategies for detecting and addressing DIF (e.g., van de Vijver & Tanzer, 2004; Zwick, 2012), but these are difficult to simulate. Therefore, we chose a DIF elimination process with an easy criterion. We varied the sample size at two levels (see Table 1) according to the initial PISA calibration sample size (OECD, 2012) and the typical sample size per state in the German National Assessment Study (Pant et al., 2013) and we varied the number of anchor items at two levels, such that the first level was typical of the PISA minor domains and the second level was somewhere between the typical numbers of anchor items in PISA's major and minor domains (Prenzel, Kobarg, Schöps, & Rönnebeck, 2012).

Calibration and Linking

Each of the three linking methods was applied in every condition. The R package TAM (Kiefer, Robitzsch, & Wu, 2015) and its functions, which are based on the marginal maximum likelihood (MML) estimation method, were used for all calibrations. Preliminary analyses showed that the results were very similar to ConQuest (Wu et al., 2007), the software used by PISA until 2015, whereby TAM was much more efficient in its use of computational resources. To estimate trends based on national data, the focal country's data at times 1 and 2 were scaled separately, that is, without using the international item parameters and linked via Stocking-Lord equating (Nat_{SL}) in one step (see Figure 1). By contrast, the trend estimation that was based on the international data was composed of three linking steps (Figure 1). First, we estimated international item parameters using the data from all countries. In line with the calibration step in past PISA cycles (2000 to 2012), we did not specify a multiple group model or make use of a latent regression model for the countries in this step because there was no interaction between item difficulty and group proficiency (cf. DeMars, 2002). Then, we estimated the national item parameters using only the data from our focal country. The person parameters were also estimated in this step, and 10 plausible values (Mislevy, Beaton, Kaplan, & Sheehan, 1992; Wu, 2005) were drawn per person. Further, we linked the focal country's national item

parameters to the international scale. Second, we repeated this procedure at time 2. Third, we linked the international item parameters at time 2 to the international scale at time 1 via SL equating and applied the resulting transformation rule to the focal country's person parameter estimates. When the linking of the focal country's item parameter estimates to the international scale (steps 1 and 2) was done via SL equating, we denote the linking method as Int_{SL}. When the focal country was linked to the international scale through fixed item parameter calibration, as was the practice in PISA 2000-2012 (OECD, 2012), we denote it as Int_{FCIP}. Note that Int_{FCIP} contains one mean-mean linking step, namely, the link between the international parameters at the two time points.

Evaluation Criteria

The summarized differences between the estimates and the generating parameters provide a good indication of the quality of parameter recovery, and therefore, an indication of the quality of the estimators. As measures of accuracy, we used the root mean squared error (RMSE) and bias. The RMSE was defined as the square root of the mean of the squared deviations of the focal country's trend estimate $\hat{\theta}_{\Delta}$ from its true population parameter $\theta_{\Delta(p)}$ across the 1,000 replications. Similarly, bias was defined as the mean deviation of the focal country's trend estimate $\hat{\theta}_{\Delta}$ from its true population parameter $\theta_{\Delta(p)}$ across the 1,000 replications. In our design, it could have been misleading to directly compare conditions with respect to RMSE or bias because the item parameters and person parameters were resampled in every condition. Thus, sampling error was confounded with the effect of condition. Therefore, we also calculated the corrected accuracy measures $\text{cRMSE}(\hat{\theta}_{\Delta})$ and $\text{cBias}(\hat{\theta}_{\Delta})$. To compute these measures, the true population trend $\theta_{\Delta(p)}$ was replaced by the true sample trend $\theta_{\Delta(s)}$, resulting in Equations 1 and 2:

$$\text{cRMSE}\left(\hat{\theta}_{\Delta}\right) = \sqrt{\frac{\sum_{i=1}^{1,000}\left(\hat{\theta}_{\Delta_i} - \theta_{\Delta(s)}\right)^2}{1,000}}, \tag{1}$$

$$\text{cBias}\left(\hat{\theta}_{\Delta}\right) = \frac{\sum_{i=1}^{1,000}\left(\hat{\theta}_{\Delta_i} - \theta_{\Delta(s)}\right)}{1,000}. \tag{2}$$

Real Data Example

To illustrate the differences between the linking methods with real data, we used the PISA 2009 and 2012 reading and mathematics assessment data (OECD, 2010, 2014a) and applied each of the three linking methods to national trends in Germany, Japan, and the United States. For the international database, we considered only the 34 OECD countries, and we drew 500 persons per country at random. The number of linking items that were used in each country between time points were 31 (math) and 43 (reading). DIF was estimated as described above by comparing the centered national anchor item parameters to the centered international anchor item parameters.

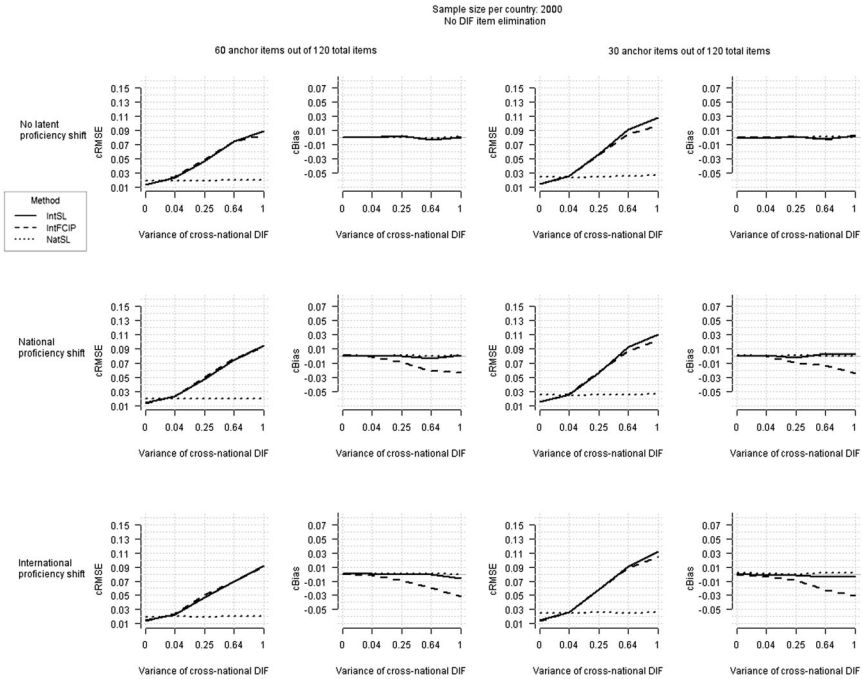

Figure 2. Interaction plots for the accuracy of the FCIP versus SL linking methods in conditions without DIF item elimination and with 2,000 persons per country with regard to cBias and cRMSE.

Results

The values of the evaluation criteria were graphed for each of the 120 combinations of DIF, latent proficiency shift, DIF item elimination, sample size, and anchor size. Figures 2 to 5 summarize the cRMSE and cBias. To test Hypothesis 1, we compared the accuracy of the trend estimates based on national item parameters to the accuracy of the trend estimates based on international item parameters, both using Stocking-Lord equating (Nat_{SL} and Int_{SL}, respectively) and using the cRMSE values as the major criterion. As can be seen in Figures 2 to 5, the results were completely in line with Hypothesis 1.

When there was no country-specific DIF, Nat_{SL} performed slightly worse than the linking methods based on international data in terms of overall error. In conditions with cross-national DIF, Nat_{SL} showed its advantages, namely, its independence from cross-national DIF, and performed better than the linking methods based on international data. This pattern of results held across all side conditions. Thus, when cross-national DIF was high, trend estimation using Int_{SL} or Int_{FCIP} was less accurate than trend estimation using Nat_{SL}, which appeared to be unaffected by cross-national DIF. Depending on the sample size per country, the interaction changed somewhat. When only 500 persons per country were considered, Nat_{SL}'s cRMSE was relatively higher than the conditions with 2,000 persons per country. Thus, in conditions with

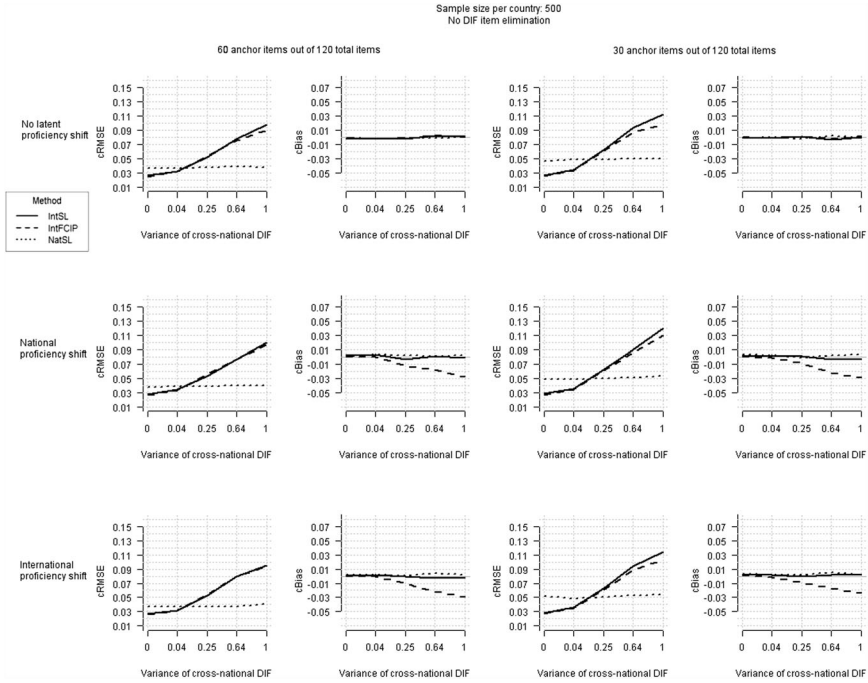

Figure 3. Interaction plots for the accuracy of the FCIP versus SL linking methods in conditions without DIF item elimination and with 500 persons per country with regard to cBias and cRMSE.

500 persons, the advantage of Nat_{SL} in terms of cRMSE could be observed in DIF conditions with high to extreme DIF, whereas in conditions with 2,000 persons these advantages showed up even in conditions with low DIF.

To test Hypothesis 2, we compared the linking methods based on international item parameters with respect to cBias and cRMSE. For all combinations of sample size and anchor size, a clear trend emerged. When a latent proficiency shift was present and when the size of the variance of the cross-national DIF was greater than a value that could be considered moderate, trend estimation using Int_{FCIP} led to bias. Accordingly, trend estimation via Int_{FCIP} was not biased per se but exhibited bias according to the amount of cross-national DIF and the proficiency shift. Trend estimation using Int_{SL} appeared to be much less biased in conditions with high DIF and a latent proficiency shift (see Figures 2 to 5). In conditions without cross-national DIF or without a latent proficiency shift, the two methods performed almost equally well with respect to cBias.

Int_{FCIP}'s bias was drastically reduced when items with an absolute DIF greater than 1 logit were excluded from the link (see Figures 4 and 5). However, the elimination of these DIF items came at the cost of an increase in the cRMSE: When the cross-national DIF was higher, there was a greater increase in the cRMSE, probably as a consequence of the reduction in anchor items. This effect held across all methods.

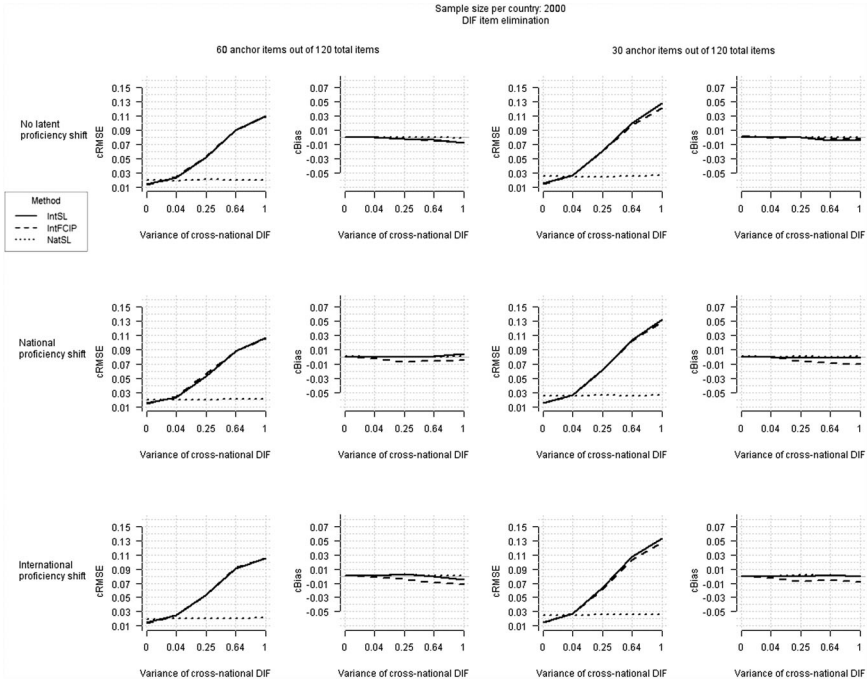

Figure 4. Interaction plots for the accuracy of the FCIP versus SL linking methods in conditions with DIF item elimination and with 2,000 persons per country with regard to cBias and cRMSE.

The RMSE summarizes an estimator's bias and its variance. Figures 6 and 7 illustrate these components for selected conditions. Regarding conditions without DIF item elimination, the trend estimator using Int_{FCIP} narrowly had the lowest variance in all conditions without cross-national DIF (Figure 6), and it had a relatively high variance in conditions with extreme DIF and with a latent proficiency shift (Figure 7).

In addition, under the latter conditions, the trend estimation using Int_{FCIP} was biased, whereas the trend estimation with Int_{SL} or Nat_{SL} appeared to be unbiased (Figure 7). The bias in trend estimation using Int_{FCIP} led to an increase in the cRMSE that could be compensated for by its relatively low variance. This resulted in near equality in cRMSE values for Int_{FCIP} and Int_{SL} across all simulated conditions, whereas Int_{FCIP} narrowly exhibited lower cRMSE values across all conditions. Thus, in conditions that are problematic for trend estimation using Int_{FCIP} (conditions with high DIF and with a latent proficiency shift), only trend estimation using national item parameters Nat_{SL} seemed to provide a good alternative because Int_{SL} exhibited a cRMSE that was similar to Int_{FCIP}'s.

In general, trend estimation using Int_{SL} or Int_{FCIP} was equally accurate with a very narrow advantage for trend estimation with Int_{FCIP} as can be seen from almost all values of cRMSE in Figures 2 to 5.

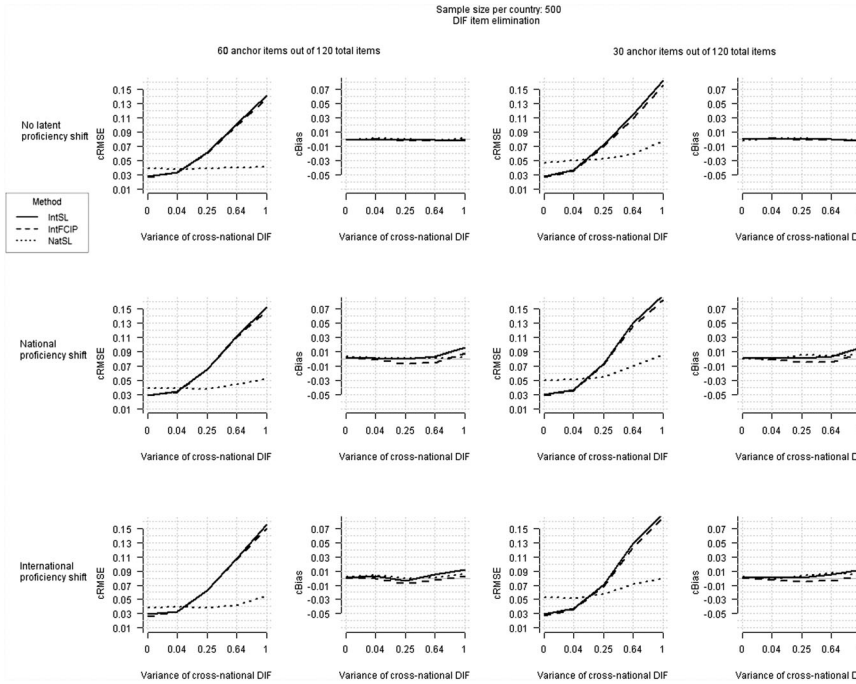

Figure 5. Interaction plots for the accuracy of the FCIP versus SL linking methods in conditions with DIF item elimination and with 500 persons per country with regard to cBias and cRMSE.

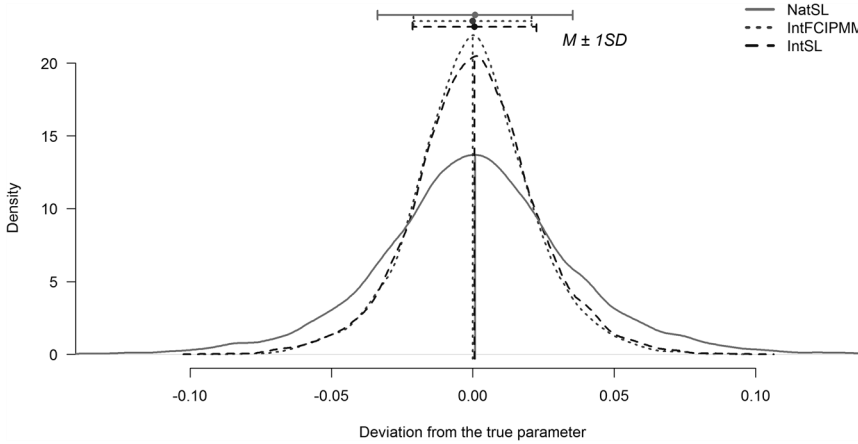

Figure 6. Comparison of the densities of the estimators' deviations from the true parameter across all conditions without DIF item elimination and without cross-national DIF.

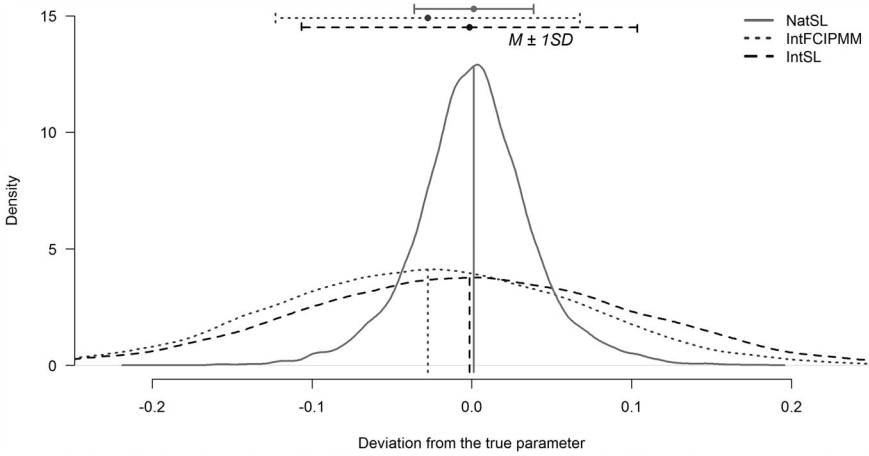

Figure 7. Comparison of the densities of the estimators' deviations from the true parameter across conditions without DIF item elimination, with a latent proficiency shift and extreme cross-national DIF ($\sigma^2_{DIF} = 1$).

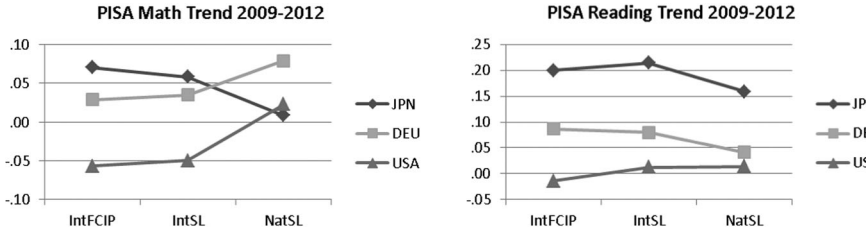

Figure 8. Comparison of trend estimates using three different linking methods for Japan, Germany, and the United States using the PISA data from the 2009 and 2012 cycles.

Regarding the side factors (DIF item elimination, sample size, and anchor size), DIF item elimination was able to reduce Int_{FCIP}'s bias to a minimum. However, the cRMSE for all methods increased when applying DIF elimination in conditions with cross-national DIF. Further, the trend estimates were generally more accurate with respect to cRMSE when 2,000 instead of 500 persons per country were simulated and when the number and percentage of anchor items were increased (Figures 2 to 5).

In the real data, we found the following DIF variances and correlations: Germany had mean estimated DIF variances of .07 (math) and .17 (reading). The respective estimated DIF correlations between 2009 and 2012 were .74 and .87. Japan had mean estimated DIF variances of .39 (math) and .31 (reading). The respective estimated DIF correlations between 2009 and 2012 were .95 and .92. The United States had mean estimated DIF variances of .11 (math) and .17 (reading). The respective estimated DIF correlations between 2009 and 2012 were .89 and .82. Further, we found differences between the linking methods (Figure 8). The differences between methods ranged from −.06 logits to .08 logits. It was largest for the U.S. math scale.

Compared with our results, it can be hypothesized that these differences occurred by chance, because they were not more than two times the roughly approximated standard error (the proxy is the RMSE when $\sigma^2_{DIF} \geq .04$, $N = 2{,}000$, anchor size $= 30$; results not shown).

Discussion

This study compared the performances of three different linking methods for national trend estimation in international large-scale assessments under various simulated conditions with different amounts of cross-national DIF. A systematic investigation of cross-national DIF in linking was essential because many methods applied to educational large-scale assessments rely on cross-national measurement invariance, and recent studies have repeatedly demonstrated that this assumption may not hold and that alternative methods should be considered (Carstensen et al., 2008; Fox & Verhagen, 2010; Gebhardt & Adams, 2007; Grisay & Monseur, 2007; Kankaraš & Moors, 2014; Kreiner & Christensen, 2014; Oliveri & von Davier, 2014; Rutkowski & Svetina, 2014). In particular, we aimed to compare the operational practice used by PISA until 2015 to two alternative methods for national trend estimation in order to identify the amounts of DIF for which each method performed best such that practitioners could get an indication of when (i.e., at what amount of DIF) it would become beneficial to switch methods.

The results of the present study indicate that in line with Hypothesis 1, linking methods that consider only national item parameters as proposed by Gebhardt and Adams (2007) may provide a superior alternative to linking based on international item parameters using the same IRT-scale transformation method, especially in cases of high DIF. In accordance with Hypothesis 2, in cases in which there was a latent proficiency shift and high DIF, the true trend was underestimated when the linking was based on international item parameters and FCIP linking, which was the operational practice in PISA until 2015 (OECD, 2012). Nevertheless, this bias did not increase Int_{FCIP}'s overall error in terms of the RMSE in comparison with the SL linking method using international data. The detection and elimination of DIF items could reduce the FCIP bias to a minimum, but it also, in turn, increased the overall error. These findings will be discussed in more detail below. Aside from the aforementioned main findings, two expected results were confirmed for all linking methods that were investigated. When estimating the true trend, the error decreased when (a) a larger sample size (2,000 persons per country as opposed to only 500) was used or (b) the number and percentage of anchor items (60 out of 120 total items as opposed to only 30 out of 120 total items) were increased.

Our investigation of Hypothesis 1 showed that using the SL linking method based on national item parameters (Nat_{SL}) led, on average, to more accurate trend estimates than using the SL linking method based on international item parameters (Int_{SL}). This advantage of Nat_{SL} in terms of a lower cRMSE appeared in conditions with moderate (or higher) cross-national DIF, which affected steps 1 and 2 of the linking based on international item parameters. Therefore, we advocate using Nat_{SL} as an alternative to Int_{SL} or Int_{FCIP} when cross-national measurement invariance is in doubt and when one is not primarily interested in international comparability.

In line with Hypothesis 2, we found differences between the linking methods that were based on international item parameters. When the variance in cross-national DIF was large according to the ETS classification (e.g., Penfield & Algina, 2006), the linking method that used fixed common international item parameters (Int_{FCIP}) was biased. This pattern held in all conditions with a latent proficiency shift, albeit eliminating items with strong DIF diminished the bias. A similar bias in estimation using FCIP when a latent proficiency shift was present has been reported before (Arai & Mayekawa, 2011; Baldwin et al., 2007; Hu et al., 2008; Kang & Petersen, 2012; Keller & Keller, 2011; Kim, 2006; Paek & Young, 2005). Our simulation showed that this effect did not occur in conditions without DIF or with negligible DIF. However, according to previous research, it seems realistic to assume at least a moderate amount of DIF between countries (Kankaraš & Moors, 2014), which is also in line with our preliminary analyses based on the PISA 2009 and 2012 data. Thus, trend estimates in international large-scale assessments based on FCIP methods may be biased, but the overall errors in FCIP and SL using international item parameters are comparable. Furthermore, the FCIP bias was reduced by DIF item elimination, but as this comes at the cost of an increase in the cRMSE, the use of the elimination procedure is somewhat limited.

With regard to practical applications, the FCIP bias that was detected was moderate, even in cases with extreme DIF. PISA reports its results on a scale with mean 500 and standard deviation 100. Expressed on this scale, Int_{FCIP}'s bias in conditions with extreme DIF did not exceed three points when a rather high latent proficiency shift of 20 points was present. However, this is the size of the standard errors that are typically reported for national trends, which are usually around three points. In conditions with average DIF variance, the bias was around one point. We also simulated only five populations, which did not affect the results. Further, we also simulated a lower latent proficiency shift (five points), which expectedly led to an even lower bias of about .5 to 1 point (results not shown) in conditions with extreme DIF.[1] We conclude that the exhibited bias should not give practitioners too many reasons to worry, taking into account that Int_{FCIP}'s total error was approximately the same as the alternative method's (Int_{SL}) total error. If one is interested in international rankings, it is, from our results, completely reasonable to use either Int_{FCIP} or Int_{SL}. Nevertheless, we encourage linking using national item parameters when cross-national DIF is suspected to be moderate or larger in size because, in extreme conditions, its sampling-error-free standard error (cRMSE) was as much as four times lower than the standard error for linking methods based on international item parameters.

Future research should investigate whether other methods, for example, the methods suggested by Oliveri and von Davier (2014) or by Fox and Verhagen (2010), could be considered alternatives that are more accurate, because our study's results provide no generalization to these methods. Further, it would be interesting to determine the nature and stability of cross-national DIF over time in different ILSAs. The instability of certain items could be taken as an indication for DIF item exclusion. Could exclusion that is enriched with information about an item's DIF over the years improve exclusion strategies? Concerning FCIP, more conditions of latent proficiency shifts (e.g., skewed distributions or other departures from normality) could be created to examine its behavior in comparison with other linking methods.

Studies could focus on sampling problems and their consequences. In addition, we used a rather simple approach to eliminate items with strong cross-national DIF. Procedures that are more sophisticated (see Zwick, 2012) might lead to more favorable results.

This study compared the performances of three methods for linking data in order to estimate trends in international large-scale assessments. Using only a few simulated conditions poses some limitations for generalizability, especially in situations in which the data are suspected to depart even further from the IRT model. Moreover, we ignored the hierarchical data structure that is common in international comparative large-scale assessments. Nevertheless, this study showed that country-specific DIF interacts with linking designs and linking methods, and such an interaction may result in different trend estimates. In our example in which we used the PISA 2009 and 2012 reading and mathematics data from three countries, we found quite different trend estimates when we applied different linking methods. This may have occurred by chance because, deduced from our simulation results, the differences between the linking methods' trend estimates did not exceed what was hypothesized to be twice their standard error. Nevertheless, we now know that in the long run properly implemented linking based on national item parameters has more favorable characteristics than linking based on international item parameters in realistic conditions with average cross-national DIF. Therefore, we recommend Nat_{SL} for researchers who are not primarily interested in international comparability when cross-national DIF is present. In addition, we identified potential problems with FCIP linking. However, the overall error of Int_{FCIP} is comparable to alternative methods using international item parameters. Thus, there seems to be no reason to discourage researchers from using this linking method, even in cases when cross-national DIF is present, where FCIP theoretically might not be appropriate.

Note

[1]The complete results can be obtained from the first author upon request.

References

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

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 Technical Report*. Technical Report NCES 2001-509. Washington, DC: National Center for Educational Statistics. Retrieved April 6, 2016, from http://nces.ed.gov/nationsreportcard/pdf/main1998/2001509.pdf

Arai, S., & Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, *38*(1), 1–16. http://doi.org/10.2333/bhmk.38.1

Arora, A., Foy, P., Martin, M. O., & Mullis, I. V. S. (Eds.). (2009). *TIMSS Advanced 2008 Technical Report*. Boston, MA: TIMSS & PIRLS International Study Center. Retrieved April 6, 2016, from http://timssandpirls.bc.edu/timss_advanced/downloads/TA08_Technical_Report.pdf

Artelt, C., & Baumert, J. (2004). Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs [Comparability of students' reading literacy performance measured with items originating from different language backgrounds]. *Zeitschrift Für Pädagogische Psychologie*, *18*(3–4), 171–185. http://doi.org/10.1024/1010-0652.18.34.171

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*, 147–162.

Baldwin, S. G., Baldwin, P., & Nering, M. L. (2007, April). *A comparison of IRT equating methods on recovering item parameters and growth in mixed-format tests*. Presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles—results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on PISA: Research outcomes of the PISA Research Conference 2009* (pp. 199–213). Dordrecht, The Netherlands: Springer. Retrieved April 6, 2016, from http://link.springer.com/chapter/10.1007/978-94-007-4458-5_12

Carstensen, C. H., Prenzel, M., & Baumert, J. (2008). Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? [Trend analyses in PISA: How did competencies in Germany develop between PISA 2000 and PISA 2006?]. *Zeitschrift Für Erziehungswissenschaft, Sonderheft*, *10*, 11–34.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: The case of Ireland and implications for international assessment practice. *Large-Scale Assessments in Education*, *2*(1), 2. http://doi.org/10.1186/2196-0739-2-2

Davidov, E., Schmidt, P., & Billiet, J. (2012). *Cross-cultural analysis: Methods and applications*. London, UK: Routledge.

De Jong, M. G., Steenkamp, J. E. M., & Fox, J. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278. http://doi.org/10.1086/518524

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, *15*(1), 15–31. http://doi.org/10.1207/S15324818AME1501_02

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 467–488). London, UK: Routledge.

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. http://doi.org/10.1111/j.1745-3992.2009.00154.x

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8*, 305–322.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*(1), 69–86. http://doi.org/10.1016/j.stueduc.2007.01.006

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

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

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement*, *32*, 311–333. http://doi.org/10.1177/0146621606292215

Huang, X. (2010). *Differential item functioning: The consequence of language, curriculum, or culture?* (Doctoral dissertation, University of California, Berkeley). Retrieved April 6, 2016, from http://escholarship.org/uc/item/1tf93776#page-3

Huggins, A. C. (2013). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement*. Advance online publication. http://doi.org/10.1177/0013164413506222

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

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*, 381–399. http://doi.org/10.1177/0022022113511297

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educational and Psychological Measurement*, *71*, 362–379. http://doi.org/10.1177/0013164410375111

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: Test analysis modules (Version 1.4-1). Retrieved from cran.r-project.org/web/packages/TAM/

Kim, D.-I., Choi, S. W., Lee, G., & Um, K. R. (2008). A comparison of the common-item and random-groups equating designs using empirical data. *International Journal of Selection and Assessment*, *16*(2), 83–92. http://doi.org/10.1111/j.1468-2389.2008.00413.x

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*, 355–381. http://doi.org/10.1111/j.1745-3984.2006.00021.x

Kolen, M. J., & Brennan, R. L. (Eds.). (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

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*, 210–231. http://doi.org/10.1007/s11336-013-9347-z

Lei, P.-W., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement*, *36*, 21–39. http://doi.org/10.1177/0146621611425171

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193. http://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160. http://doi.org/10.1111/j.1745-3984.1977.tb00033.x

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*, 133–161. http://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*, 545–563. http://doi.org/10.1007/s10763-010-9206-7

OECD. (2010). *PISA 2009 results: What Students know and can do: Student performance in reading, mathematics and science (Volume I)*. Paris, France: OECD.

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

OECD. (2014a). *PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science (Volume I, revised edition, February 2014)*. Paris, France: OECD.

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

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and Assessment Modeling*, *53*, 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. http://doi.org/10.1080/15305058.2013.825265

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Los Angeles, CA: Sage.

Paek, I., & Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education*, *18*, 199–215. http://doi.org/10.1207/s15324818ame1802_4

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *The IQB national assessment study 2012: Competencies in mathematics and the sciences at the end of secondary level I*. Münster, Germany: Waxmann. Retrieved April 6, 2016, from http://www.iqb.hu-berlin.de/laendervergleich/laendervergleich/lv2012/Bericht/IQB_NationalAsse.pdf

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*, 295–312. http://doi.org/10.1111/j.1745-3984.2006.00018.x

Prenzel, M., Kobarg, M., Schöps, K., & Rönnebeck, S. (2012). *Research on PISA: Research outcomes of the PISA Research Conference 2009*. Dordrecht, The Netherlands: Springer Science & Business Media.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. http://doi.org/10.1177/0013164413498257

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227–253. http://doi.org/10.1080/08957340701301207

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119–135. http://doi.org/10.1016/j.erap.2003.12.004

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3*(3), 115–124. http://doi.org/10.1027/1614-2241.3.3.115

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

Wu, M. L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. http://doi.org/10.1111/j.1745-3992.2010.00190.x

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest Version 2: Generalised item response modelling software [computer program]. Camberwell: Australian Council for Educational Research.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i–30. http://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Authors

KAROLINE SACHSE is a Researcher at the Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin; karoline.sachse@iqb.hu-berlin.de. Her primary research interests include trend estimation in large-scale assessments.

ALEXANDER ROPPELT is a Researcher at the Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin; ar.research@familie-roppelt.de. His primary research interests include mathematics education and methodological aspects of large-scale assessments.

NICOLE HAAG is a Researcher at the Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin; nicole.haag@iqb.hu-berlin.de. Her primary research interests include differential item functioning and effects of test items' linguistic complexity on the performance of second language learners.