The Benefits of Fixed Item Parameter Calibration for Parameter Accuracy in Small Sample Situations in Large-scale Assessments

Christoph König

Goethe University Frankfurt am Main, Germany

Lale Khorramdel and Kentaro Yamamoto

Educational Testing Service, Princeton, New Jersey

Andreas Frey

Goethe University Frankfurt am Main, Germany

Centre for Educational Measurement at the University of Oslo, Norway

Author Note

Christoph König, Department of Educational Psychology, Goethe University Frankfurt am Main, Germany; Lale Khorramdel and Kentaro Yamamoto, Educational Testing Service, Princeton New Jersey; Andreas Frey, Department of Educational Psychology, Goethe University Frankfurt am Main, Germany, Centre for Educational Measurement at the University of Oslo, Norway.

Lale Khorramdel is now at National Board of Medical Examiners, Philadelphia, Pennsylvania.

Correspondence concerning this article should be addressed to Christoph König, Goethe University Frankfurt am Main, Theodor-W.-Adorno Platz 6, 60629 Frankfurt am Main, Germany. E-Mail: koenig@psych.uni-frankfurt.de

The Benefits of Fixed Item Parameter Calibration for Parameter Accuracy in Small Sample Situations in Large-Scale Assessments

Over the years, international large-scale assessments such as the Programme for International Student Assessment (PISA), developed by the Organisation for Economic Co-operation and Development (OECD), have become integral parts of the field of educational science. They are a major source of information for both educational research and practice, and have considerable impact on national educational governance (Fischman et al., 2019). Some countries, such as Germany or Austria, for instance, established educational standards and introduced a nationwide assessment system after the so-called "PISA shock" of its first administration, in reference to their students' unexpectedly low performance compared to those from other countries (Ertl, 2006; Waldow, 2009).

The impact on national educational governance is not without debate. A large body of research documents challenges related to, for instance, the sampling design, data quality, and appropriateness of the underlying item response theory (IRT) models (Hopfenbeck et al., 2018). Frey and Hartig (2020) identify five current methodological challenges, such as completing the introduction of computerized adaptive testing, that should be addressed to allow large-scale assessments to continue to provide highly useful information on educational outcomes in the future. Rutkowski (2018) argues that to retain and increase the utility of international large-scale assessments for intended stakeholders, it is necessary to increase flexibility in design and implementation—in other words, quickly incorporate features that represent the most current understanding of assessment frameworks, psychometric models, and delivery platform capabilities.

To accomplish this, most large-scale assessments have field trials where the practicability of new features is tested. The number of new features that can be tested, however,

is limited by the sample size of these trials. Moreover, since the sample is also used for an initial IRT scaling, not all of it can be used for testing new features. Fixed item parameter calibration (FIPC; e.g., Kim, 2006), one of several calibration methods useful in situations when assessments include both old items (those used in prior assessment rounds or cycles; trend items in PISA) and new ones (those implemented for the first time), may be promising when sample size is critical. Especially because FIPC, in contrast to other calibration methods, allows introducing prior information into the calibration. Consequently, the current study investigates if using FIPC reduces the sample required for an accurate initial IRT scaling of field-trial assessment data. The smaller the proportion of the sample required, the larger the proportion available for testing new features to implement in the main survey. This, in turn, contributes to increased flexibility of large-scale assessments in design and implementation.

Current Practice for PISA's Main Survey

PISA is an international large-scale assessment testing the skills and knowledge of 15-year-old students in three core domains (Mathematics, Reading, and Science). Its main survey is administered every three years, with each rotating as the major domain. In 2015, PISA moved from a paper- to a computer-based assessment, and in 2018, a multistage adaptive testing design was introduced for the Reading assessment (Yamamoto, Shin, & Khorramdel, 2018). The target sample size of PISA is 6,300 students per country, obtained with a complex two-stage stratified sampling design involving a random sample of schools in the first stage, and a random sample of students in the second (Organisation for Economic Co-operation and Development [OECD], 2017a; OECD, 2017b).

The main surveys are administered to students following a balanced incomplete block design (BIBD; e.g., Frey et al., 2009). Thus, not all students respond to all items. The advantage of the BIBD is that more items can be included in the assessment, providing broader construct coverage without increasing the burden for individual students (Gonzalez & Rutkowski, 2010).

Since 2015, the IRT scaling of the assessment data has been based on a concurrent calibration with country-by-language groups as the grouping variable and the assumption of partial invariance. For the scaling of trend items (i.e., items developed and administered in PISA assessments prior to 2015), a hybrid model is used that combines the Rasch (Rasch, 1960) and partial credit models (PCM; Masters, 1982) with the two-parameter logistic model (2PLM; Birnbaum, 1968) and generalized partial credit model (GPCM; Muraki, 1992) for dichotomous and polytomous items, respectively (OECD, 2017c). For new items developed for the 2015 assessment and later, only the 2PLM and GPCM are used. To examine the fit of the hybrid model (compared to the Rasch model and PCM), von Davier et al. (2019) conducted a recalibration using PISA data from 2000 to 2012. It showed that the more general (hybrid) model yielded a better global model-data fit than the Rasch model and PCM, which were historically used in PISA before the 2015 assessment. Moreover, this new modeling approach also resolved many of the item-level misfit issues that appeared in the Rasch model.

Differential item functioning (DIF) is treated with a partial invariance approach that assumes invariance of item parameter estimates across multiple groups for most items. More precisely, international item parameters are estimated for most items (i.e., the same item parameters are estimated across all groups), while for a subset of items, country-specific unique item parameters are allowed in cases where DIF occurs. This approach allows the comparability of item parameter estimates and data across different countries and languages while simultaneously accounting for item misfit (for more details, see OECD, 2017c).

PISA data consist of responses to dichotomous and polytomous trend and new items, which can be further divided into multiple-choice and open-ended response items. The latter are coded by human raters and a machine-supported coding system, which was introduced for short responses (Yamamoto, He, Shin, & von Davier, 2018). In addition to trend items from previous PISA assessments, new items are developed between assessments to ensure adequate

construct coverage based on changing frameworks. They are developed only for the major domain of the respective PISA assessment. For example, for the 2015 main survey, 99 new items were developed for Science because it was the major domain.

Trend items are used for linking multiple PISA assessments and to report trends on a common scale across assessments (OECD, 2017c). Prior to PISA 2015, the data in each assessment were calibrated separately and subsequently equated. This changed in PISA 2015, when a multiple group (concurrent) FIPC replaced the separate calibration approach. First, common item parameters across different countries, languages, and assessments were estimated with a multiple group concurrent calibration of data from the PISA 2006, 2009, 2012, and 2015 cycles. Second, the resulting item parameter estimates have been used to fix the trend items in PISA 2018 to link the 2018 assessment to past PISA assessments. The trend item parameters are updated in every main survey to account for item-by-country and item-by-language interactions (note that parameters are only updated in a few cases where necessary). This involves examining their fit; in case of misfit, unique (i.e., country-specific) item parameters are estimated. The updated estimates are then used to fix the trend items in the subsequent field trial (for estimating preliminary item parameters for newly developed items on the trend scale to select items for the main survey), and the related main survey (for estimating the final item parameters).

Current Practice for PISA's Field Trial

PISA field trials are conducted a year before the respective main surveys (e.g., the field trial for the 2015 main survey was conducted in 2014). These field trials have a dual purpose. First, they are used to estimate preliminary item parameters and ability estimates in an initial IRT scaling. The assessments administered in the field trials typically consist of a larger number of new items (OECD, 2017c). Scaling results are used to evaluate the quality of these items and to select among them for the main survey. To establish a link to the international

scale (i.e., to past assessment cycles), the parameters for trend items are fixed to the estimates obtained in the last main survey (OECD, 2017c). The parameters for new items are estimated freely but scaled together with trend items on a common scale. In PISA 2015, trend item parameters were estimated by a calibration including the 2000–2015 data (von Davier et al., 2019); in PISA 2018, trend item parameters were fixed to estimates obtained from the PISA 2015 main survey calibration.

Second, they are used to evaluate new psychometric and assessment features (related to the assessment design, mode of assessment, or nature of the assessed construct) regarding their feasibility for the main survey. New psychometric and assessment features, however, need to be thoroughly evaluated; to ascertain that a new feature can be implemented successfully in the main survey, multiple research studies in the field trial are needed without decreasing the comparability of item parameter estimates or harming the measurement of trend over time (both are important goals of international large-scale assessments).

The desire to keep the main survey up to date with developments in the field of psychometrics and to meet stakeholders' needs regarding the comparability of scales and stable trend measures places large demands on PISA field trials, since their samples are considerably smaller than those of the main surveys. A school-based simple random sample of 25 schools with 78 students per school yields a target sample size of 1,950 students per country (OECD, 2017c). This is problematic, because the underlying IRT models require relatively large sample sizes for an accurate item calibration and scaling (the recommended sample size for single-group models is $N = 500$, de Ayala, 2014; for multigroup models, however, $N = 500$ is considered a small sample size, Kim & Kolen, 2019). Moreover, Mazzeo and von Davier (2014) show that the country-specific effective sample size of the PISA main survey, after taking into account the sampling and booklet design, actually ranges from $n_e = 250$ to $n_e = 750$ responses per item per country. Consequently, two questions arise. First, how can the field

trial incorporate more comprehensive tests of new assessment features without an increased sample size? Second, what proportion of the available field-trial sample is critical to retain the accuracy of the initial item calibration and scaling of the assessment data in the field trial? As stated earlier, FIPC (e.g. Kim, 2006; Kang & Petersen, 2012), one of several calibration methods that are useful in situations where assessments consist of old and new items, may be promising when sample size is critical.

Purpose of the Study and Research Questions

By fixing the parameters of trend items to their previously obtained estimates, FIPC introduces prior information into the calibration process. This can be especially helpful for improving the accuracy of item parameter estimates in small samples. The utility of this prior information and, more specifically, of FIPC for small-sample item calibration in the context of international large-scale assessments thus far has not been studied extensively. Recently, Kim and Kolen (2019) showed that FIPC applied to multigroup settings performs better than Stocking-Lord equating (a separate calibration followed by equating) in recovering the underlying ability distributions and parameter estimates of new items. They did not focus, however, on the impact of different types of prior information introduced into the calibration. Other research on FIPC indicates that it may yield biased model parameters under certain conditions. Bias depends on the sample size (Hanson & Béguin, 2002; Kang & Petersen, 2012), the number of items with parameters available from previous calibrations (e.g., Arai & Mayekawa, 2011; Kim et al., 2018), the amount of cross-national DIF (Sachse et al., 2018), and shifts in the latent ability distributions across assessments (e.g., Baldwin et al., 2007; Keller et al., 2007). Keller and Keller (2011; 2015), however, showed that FIPC works best for complex changes in the latent ability distributions and in cases where the content of the assessment changes. Zhao and Hambleton (2017) showed that FIPC was robust against ability shifts across two adjacent assessments. Moreover, Paek and Young (2005) and Kim (2006)

showed that the performance of FIPC depends on the implementation of the marginal maximum likelihood (MML) estimation routine in IRT software packages: Performance of FIPC is best when the software uses multiple updates of the prior latent ability distribution combined with multiple expectation-maximization (EM) cycles (the MWU-MEM implementation; Kim, 2006). Although sample size has been considered in previous studies on FIPC (e.g., Hanson & Béguin, 2002; Kang & Petersen, 2012), sample sizes smaller than $N = 500$ have not been examined. Moreover, the performance of FIPC in multigroup settings with more than two groups and complex test designs and sampling frames has not been considered.

Therefore, the purpose of this study is to investigate the impact of FIPC on the accuracy of item parameter estimates in small samples using real PISA 2015 Science assessment data. More precisely, the purpose is to determine how accurate the estimation of item parameters for new items in different sampling conditions is if we can utilize different amounts of prior information (i.e. prior information from multiple assessments, prior information from only one assessment, and no prior information) by fixing trend item parameters in FIPC. The first and second research questions relate to the outcome of FIPC in small samples: When using FIPC in increasingly smaller samples, (1) how does the accuracy (in terms of bias and standard error) of the parameter estimates of the new Science items (human- and machine-coded) change, and (2) how does the fit (in terms of mean deviation and root mean squared deviation) of the trend and new Science items (human- and machine-coded) change?

The third and fourth research questions relate to the input and consequence of FIPC in small samples: When using FIPC in increasingly smaller samples, (3) what is the impact of fixing the trend Science items to different estimates (i.e., introducing different amounts of prior information into the calibration) on the accuracy of the parameters of the new Science items, and (4) what is the critical sample size that still provides accurate item parameter estimates?

For a complete picture of the performance of FIPC in small samples, one final research question relates to the estimation of the country-specific latent trait distributions: (5) what is the impact of FIPC on the means and standard deviations of the country-specific latent trait distributions? Answers to these questions provide an indication how FIPC contributes to PISA's flexibility; in other words, how much of the current field-trial sample size is required for the initial item calibration and scaling, and how much can be used to test new features for implementation in the main survey.

Method

Data

This study utilized empirical data from the PISA 2015 main survey, when Science was the major domain. For the current study, 10 countries (language group in parentheses) that took PISA as computer-based assessment were selected: Australia (English), Denmark (Danish), Finland (Finnish), France (French), Germany (German), Italy (Italian), Japan (Japanese), Malaysia (Malay), Taipei (Chinese), and the United States (English). These countries were chosen due to their diversity in culture as well as languages: both alphabetic-based languages (European, such as Danish, English, Finnish, French, German, and Italian) and character-based languages (Asian such as Chinese, Japanese, and Malay) are represented. In total, the sample consisted of $N = 76{,}722$ students. This is referred to as the "study sample" in the remainder of this paper.

Procedure

Design. First, the design consisted of four different sampling conditions. Subsamples of size $n = 125, 250, 500, 1{,}000$ students were drawn from the samples for each selected country using a school-based bootstrapping approach (outlined in the following section), resulting in datasets of $N = 1{,}250, 2{,}500, 5{,}000, 10{,}000$ students. Second, three types of parameter constraints were used. Each type of parameter constraints introduces a different

amount of prior information into the calibration (a multigroup concurrent calibration based on the 2PLM and GPCM with partial invariance assumption, i.e. common item parameters across countries):

a. FIPC-PISA (FIPC-P): Trend Science items are fixed to their final estimates obtained from the PISA 2015 main survey. Note that by fixing the trend item parameters to the 2015 main survey estimates, prior information from multiple assessments (PISA 2006 to 2015) is introduced into the analysis.
b. FIPC-Study (FIPC-S): Trend Science items are fixed to their estimates obtained from a multiple group concurrent calibration with common item parameters across countries based on the study sample data ($N = 76{,}722$) in each condition. Note that by fixing the trend item parameters to the study sample estimates, prior information from one assessment (PISA 2015) is introduced into the analysis. Thus, the amount of prior information is smaller compared to FIPC-P, but larger compared to CC-S, where no prior information is utilized.
c. CC-Study (CC-S): Trend and new Science items are freely estimated in each condition; that is, not fixed to any values from a previous calibration. Thus, no prior information is utilized in the calibration. Trend and new Science items are on the base scale by constraining the item parameters to be equal across countries (the partial invariance assumption mentioned above).

Hence, the IRT scaling applied in this study mimics scaling approaches in large-scale assessments in general, and PISA in particular, but uses different constraints on the trend items as described above. To evaluate the performance of FIPC adequately, sample sizes were chosen to be well below the recommended sample size for both the 2PLM and the GPCM, as well as below the typical country-specific target sample size of the PISA field trial. Moreover, the trend Science item parameter estimates in FIPC-P are based on a larger dataset (i.e., more

information is utilized for the parameter estimation) than in FIPC-S. Additionally, FIPC-P accounts for DIF by including country-specific unique trend Science item parameter estimates, while FIPC-S does not.

Data sampling. Important goals of the data sampling process were a) retaining the hierarchical structure of the PISA 2015 data (students nested in schools) and mimicking the sampling approach in the PISA field trial (a simple school-based convenience sample). Therefore, a school-based bootstrapping approach was applied to the study sample to sample the datasets for the analyses. Thus, 1.6%, 3.2%, 6.5%, and 13.2% of schools, respectively, were drawn with replacement from the country-specific school samples. These percentages were necessary to obtain the desired subsample sizes of $n = 125, 250, 500, 1{,}000$. This ensured that only whole schools were included in the country-specific subsamples while adequately considering the size of the countries' school samples. For example, from the 177 schools in the study sample of the United States, 3.2% were drawn with replacement to obtain a country-specific subsample size of $n \sim 500$. This was repeated for each of the other nine countries, resulting in a dataset of approximately $5{,}000$ students. To ensure that the results were generalizable beyond a single random sample, this sampling process was repeated 100 times, resulting in 100 datasets of each size $N \sim 1{,}250, 2{,}500, 5{,}000, 10{,}000$ (400 datasets in total). The resampling was done in R (R Core Team, 2018) with a custom function written by the authors. The resulting average country-specific sample sizes $\bar{n}$, as well as the average sizes of the sampled datasets $\bar{N}$, are summarized in Table 1.

Analysis. The bootstrapped datasets consisted of responses to 184 computer-based items of the 2015 Science assessment. Eighty-five items were trend Science (i.e., items already calibrated in previous PISA assessments), and 99 were new Science (developed for the 2015 Science assessment). Multigroup IRT models based on the 2PLM and GPCM with country as the grouping variable were fitted to each of the 400 datasets. This is the current operational

scaling approach used in PISA. Model fitting was done with the software *mdltm* (von Davier, 2005). The *mdltm* software allows the application of the mixture general diagnostic modeling framework (MGDM), which includes multigroup IRT models based on the 2PL and GPCM as special cases (von Davier, 2010). It provides MML estimates obtained using customary EM methods. It was designed to handle large datasets as well as complex test and sampling designs. It allows the estimation of a number of different latent variable models, includes different constraints for parameter estimation, and provides different model and item fit statistics as well as methods for proficiency estimation. In addition, it can handle missing data by design and nonresponse, as well as multiple populations and weights to account for complex sampling (for a detailed description, see Khorramdel et al., 2019).

Table 1

Summary of the Data Sampling Process, Averaged Over 100 Replications

	Proportion of Schools				
Country	100%	1.6%	3.2%	6.5%	13.2%
Australia	7,939	123	264	515	1,046
Denmark	7,760	116	248	503	1,024
Finland	7,620	140	230	466	996
France	7,375	127	253	501	1,042
Germany	7,939	120	246	521	1,022
Italy	7,878	134	259	508	1,021
Japan	7,939	120	240	524	1,042
Malaysia	6,447	118	218	414	835
Taipei	7,939	110	259	516	1,038
United States	7,886	131	265	543	1,016
Average Subsample ($\bar{n}$)	7672.2	123.9	248.2	501.1	1008.2
Average dataset ($\bar{N}$)	76,722	1,239	2,482	5,011	10,082

For CC-S, the sum of item difficulties was set to zero, and the mean of the item discriminations was set to one so that the distribution of the latent trait was freely estimated in relation to the item locations. For both FIPC-P and FIPC-S, these constraints were not needed, because fixing the parameter estimates of the trend Science items set the scale for the new

Science items. For the latent trait distributions, 41 quadrature points θ_t (ranging from -5 to 5) were specified and set equal across countries. Senate weights were used to provide equal contributions of the countries to the calibration results and to mimic the usual estimation procedure in PISA. Senate weights sum up to the same value for each country. They make the population of each country to be, for instance, $N = 1{,}950$, to ensure an equal contribution for each of the countries in the analysis (OECD, 2017d).

Dependent measures. Bias in item parameter estimates was indicated by $Bias_i = \xi_i - \xi_i^{2015}$, where $\xi_i = \alpha_i, \beta_i$ was the difference between the parameter estimates of the new Science items resulting from the different constraints on the trend Science items and their corresponding estimates obtained from the PISA 2015 Science assessment (von Davier et al., 2019). These estimates were chosen as baseline because IRT scaling conducted in 2015 used data from multiple previous PISA assessments (from 2006 to 2015) and are considered the most accurate to date (OECD, 2017a). For each condition, bias was averaged over replications r ($r = 1, \ldots, 100$) to obtain summary indices for each item, that is $Bias_i = \frac{1}{r}\sum_r \xi_{ri} - \xi_i^{2015}$. To identify differences in performance between FIPC-P, FIPC-S, and CC-S differences in the estimated marginal mean bias were evaluated with the emmeans R-package (Lenth, 2020).

Precision of the item parameter estimates was indicated by their average standard errors. In each replication, for each parameter type (discrimination, difficulty, and step parameters) and item, the associated standard error of the parameter estimate was calculated. The resulting standard errors were squared to obtain the error variances for each parameter type and item, which were then averaged over replications and items. The average standard errors reported in the results were obtained by taking the square root of the averaged error variances.

To assess the item fit of the new and trend Science items, the mean deviation ($MD_i = \int (P_{0i}(\theta) - P_{ei}(\theta))\, f(\theta) d\theta$), and the root mean square deviation ($RMSD_i = \sqrt{\int (P_{0i}(\theta) - P_{ei}(\theta))^2\, f(\theta) d\theta}$; e.g., Buchholz & Hartig, 2019) were calculated as item fit

indices or measures. Both are used as item fit measures in the operational PISA analyses and are, therefore, utilized in our study as well. $P_{0i}(\theta) - P_{ei}(\theta)$ describes the deviation of the observed item characteristic curve from its expected counterpart for a given ability level θ, and $f(\theta)$ is the density of ability distribution at this ability level. For both MD and RMSD, numerical integration of θ is based on summation over the finite grid of quadrature points θ_t mentioned above. MD and RMSD quantify the magnitude and direction of deviations in the observed data from the estimated item characteristic curves and provide complementary information. MD values close to zero indicate there are no discrepancies between the observed and estimated item characteristic curves, that is, perfect item fit (Yamamoto, Khorramdel, & Shin, 2018). The MD is more sensitive to deviations of observed item difficulties than the RMSD. The RMSD is more sensitive to the deviations of both the item difficulties and discriminations (OECD, 2017c). While there are no general cutoff values for both MD and RMSD, in accordance with common practice in PISA and the Programme for the International Assessment of Adult Competencies (OECD, 2017c; Yamamoto et al., 2018), MD values between ± 0.15 and RMSD values smaller than 0.15 indicated acceptable item fit in the current study. All dependent measures were averaged over replications to obtain summary indices for each item. Similar to the bias of the item parameter estimates, MD and RMSD values were averaged over replications r to obtain summary indices for each item, that is, $MD_i = \frac{1}{r}\sum_r MD_{ri}$, and $RMSD_i = \frac{1}{r}\sum_r RMSD_{ri}$. The means and standard deviations of the country-specific latent trait distributions resulting from the different calibration methods were also averaged over replications, i.e. $M_{Country} = \frac{1}{r}\sum_r M_{Countryr}$ and $SD_{Country} = \frac{1}{r}\sum_r SD_{Countryr}$.

Results

Parameter Constraints Crucial for Parameter Recovery

This section illustrates the recovery of the parameter estimates of the new Science items overall (Figure 1), and broken down into bias in the human- and machine-coded new Science items (Figure 2).

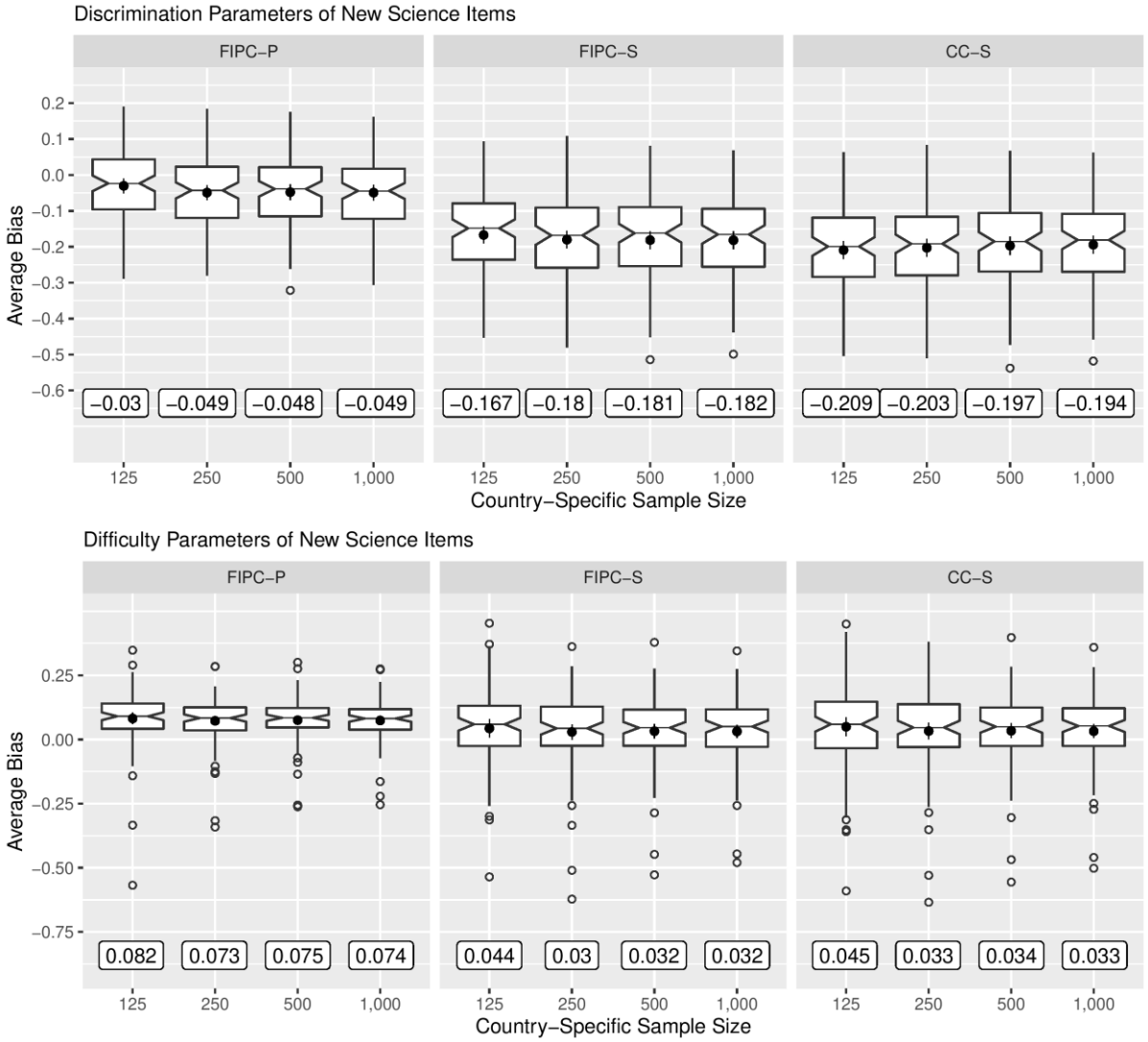

Figure 1. Average bias in item discriminations and difficulties of new Science items across sample sizes and type of parameter constraints, averaged across replications. Note: The black dot is the mean bias; its value is depicted in the boxes below the boxplots. SEs of the means range from 0.009 to 0.021.

The emerging pattern of results is characterized as follows. Item discriminations, on the one hand, are underestimated in the case of CC-S (no prior information introduced) and FIPC-S (trend item parameters fixed to their study sample parameter estimates, introducing prior

information from one assessment) across all subsample sizes. Underestimation is smaller for FIPC-P, which fixed the trend item parameters to their PISA 2015 estimates (introducing prior information from multiple PISA assessments). Item difficulties, on the other hand, are marginally overestimated for both FIPC alternatives. Compared to the parameter recovery of CC-S, overestimation is slightly larger in the case of FIPC-P.

This general pattern also applies to the parameter recovery of the machine- and human-coded new Science items (see Figure 2). Compared to the overall results, however, for the human-coded new Science items, the underestimation of the item discriminations and the overestimation of the item difficulties is more distinct. Bias in the machine-coded new Science items is more similar to the overall results.

Overall, FIPC-P especially exhibits differences from CC-S in terms of parameter recovery. FIPC-P has advantages regarding the item discrimination parameters over CC-S, and marginal disadvantages regarding the item difficulties. FIPC-S performs similar to CC-S regarding both item discrimination and difficulty parameters. This result is confirmed by contrasting the estimated marginal mean bias in the item parameter estimates across parameter constraints (see the graphical summaries in Figures S1, S2, and S3 in the supporting information). Thus, care has to be taken regarding the estimates of the trend items. The prior information they introduce are a more crucial factor for item parameter recovery than sample size. On average, bias in item parameter estimates does not change markedly across sample sizes, even in the smallest sample size.

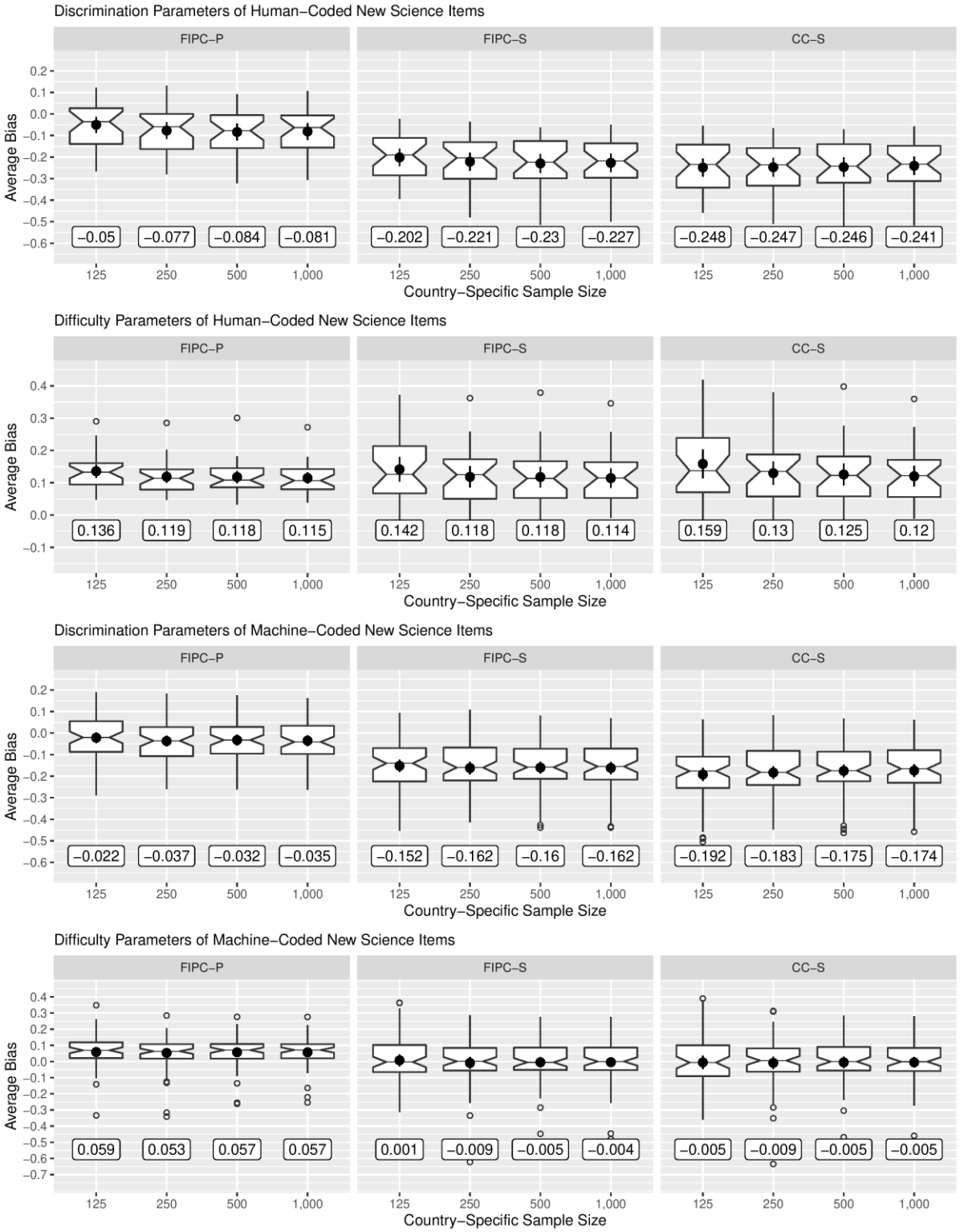

Figure 2. Average bias in item parameter estimates of human and machine-coded new Science items across subsamples and type of parameter constraints, averaged over replications. Note: The black dot is the mean average bias; its value is depicted in the boxes below the boxplots. SEs of the means range from 0.009 to 0.026.

FIPC Does Not Compensate for Loss in Precision of Item Parameter Estimates

Figure 3 illustrates the changes in precision of the item parameter estimates for the new Science items across sample sizes. Because the standard errors are virtually indistinguishable across types of parameter constraints, only results for FIPC-P are shown.

In general, as the sample size decreases, the standard errors of the item parameter estimates increase. This is the case for all types of item parameters (discrimination, difficulty, and step parameters). The largest average standard error is consistently associated with the smallest subsample sizes ($n = 125$): $M_{SE_\alpha} = 0.219$ for discrimination, $M_{SE_\beta} = 0.197$ for difficulty, and $M_{SE_\delta} = 0.241$ for step parameters. Compared to $n = 1{,}000$, the average standard error is more than twice as large. Performance of FIPC is similar to CC-S except for FIPC-P in the case of item discrimination, where the standard errors are slightly higher compared to FIPC-S or CC-S.

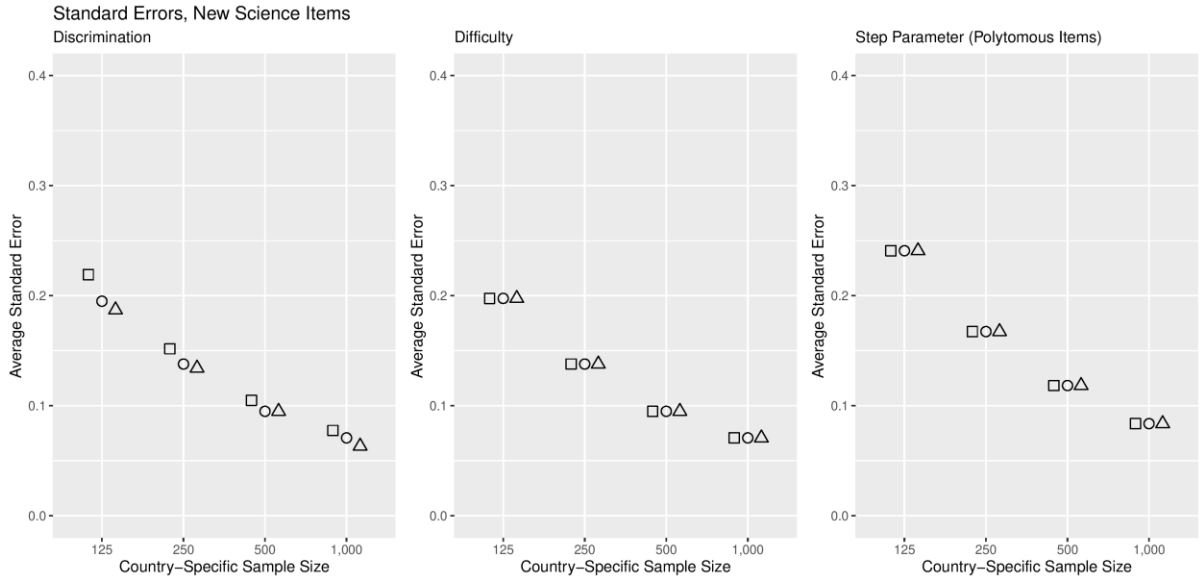

Figure 3. Average standard errors of the parameter estimates of the new Science items across types of parameter constraints and subsample sizes, averaged over replications and items. □ = FIPC-P; ○ = FIPC-S; △ = CC-S.

This pattern of increasing average standard errors remains the same for machine- and human-coded items (see Table S1 in the supporting information).

Overall, the change in precision of the item parameter estimates in terms of their average standard errors is largely independent of the type of parameter constraints. On the one hand, neither FIPC-P nor FIPC-S compensate for the loss of precision; on the other, the performance of FIPC-S is similar to CC-S across all subsample sizes.

Item Fit Measures Remain Acceptable Until $n = 250$

Table 2 illustrates changes in RMSD values of the new and trend Science items.

Table 2

Average RMSD of Science Items across Types of Parameter Constraints and Subsample Sizes

	New Science Items			Trend Science Items		
n	FIPC-P	FIPC-S	CC-S	FIPC-P	FIPC-S	CC-S
125	0.162 (0.003)	0.162 (0.003)	0.162 (0.003)	0.196 (0.003)	0.206 (0.003)	0.201 (0.003)
250	0.126 (0.003)	0.126 (0.003)	0.126 (0.003)	0.148 (0.002)	0.159 (0.002)	0.156 (0.002)
500	0.100 (0.003)	0.100 (0.003)	0.100 (0.003)	0.112 (0.002)	0.124 (0.002)	0.122 (0.002)
1,000	0.083 (0.003)	0.083 (0.003)	0.083 (0.003)	0.085 (0.001)	0.100 (0.002)	0.099 (0.002)

Note. n = sample size per country; RMSD = root mean square deviation; standard errors in parentheses. Acceptable item fit indicated by RMSD values $< .15$.

The average RMSD of new and trend Science items increases linearly across sample sizes. In the case of the new Science items, the average RMSD remains well below the threshold of 0.15 until a subsample size of $n = 250$. In the smallest subsample size ($n = 125$), the average RMSD is closer to the defined threshold. This pattern is similar for machine- and human-coded items. For machine-coded items, the average RMSD in subsamples of $n = 125$ and $n = 250$ is 0.161 ($SE = 0.004$) and 0.124 ($SE = 0.004$), respectively. For human-coded items, the average RMSD in subsamples of $n = 125$ and $n = 250$ is 0.164 ($SE = 0.007$) and 0.129 ($SE = 0.006$), respectively. For the trend Science items, there is an

increase in average RMSD compared to the average RMSD of the new Science items. Moreover, FIPC-P exhibits a smaller average RMSD than the other calibration methods.

There are no differences in average MD across subsample sizes or item types. In all cases, the average MD is close to zero, indicating no discrepancies in the observed and expected item characteristic curves (see Supporting Information, Table S2).

Overall, item fit indices of new, trend, machine-, and human-coded Science items are unacceptable in the smallest subsample size $n = 125$, regardless of type of parameter constraints. Both FIPC methods yield item fit indices comparable to CC-S, with a small advantage for FIPC-P in terms of the RMSD of the trend Science items.

Parameter Constraints Are Vital for Recovery of Latent Trait Distributions

Similar to the bias of the item parameter estimates, the performance of FIPC in terms of estimated means and standard deviations of the country-specific latent trait distributions depends on the estimates to which the trend Science item parameters are fixed, i.e. the prior information introduced into the calibration. For FIPC-P, the following pattern emerges: For countries with a low mean ability, the mean of the latent trait distribution is higher compared to CC-S. At the same time, the standard deviation is smaller. The differences in the estimated means disappear, however, as the mean ability of the respective country increases. For the standard deviations the differences remain. In the case of FIPC-S, there are no differences in the means and standard deviations of the country-specific latent trait distributions compared to CC-S. Figure 4 illustrates this pattern by reference to four countries with low (Malaysia; MYS), medium (Australia; AUS, and United States; USA), and high (Japan; JPN) mean ability.

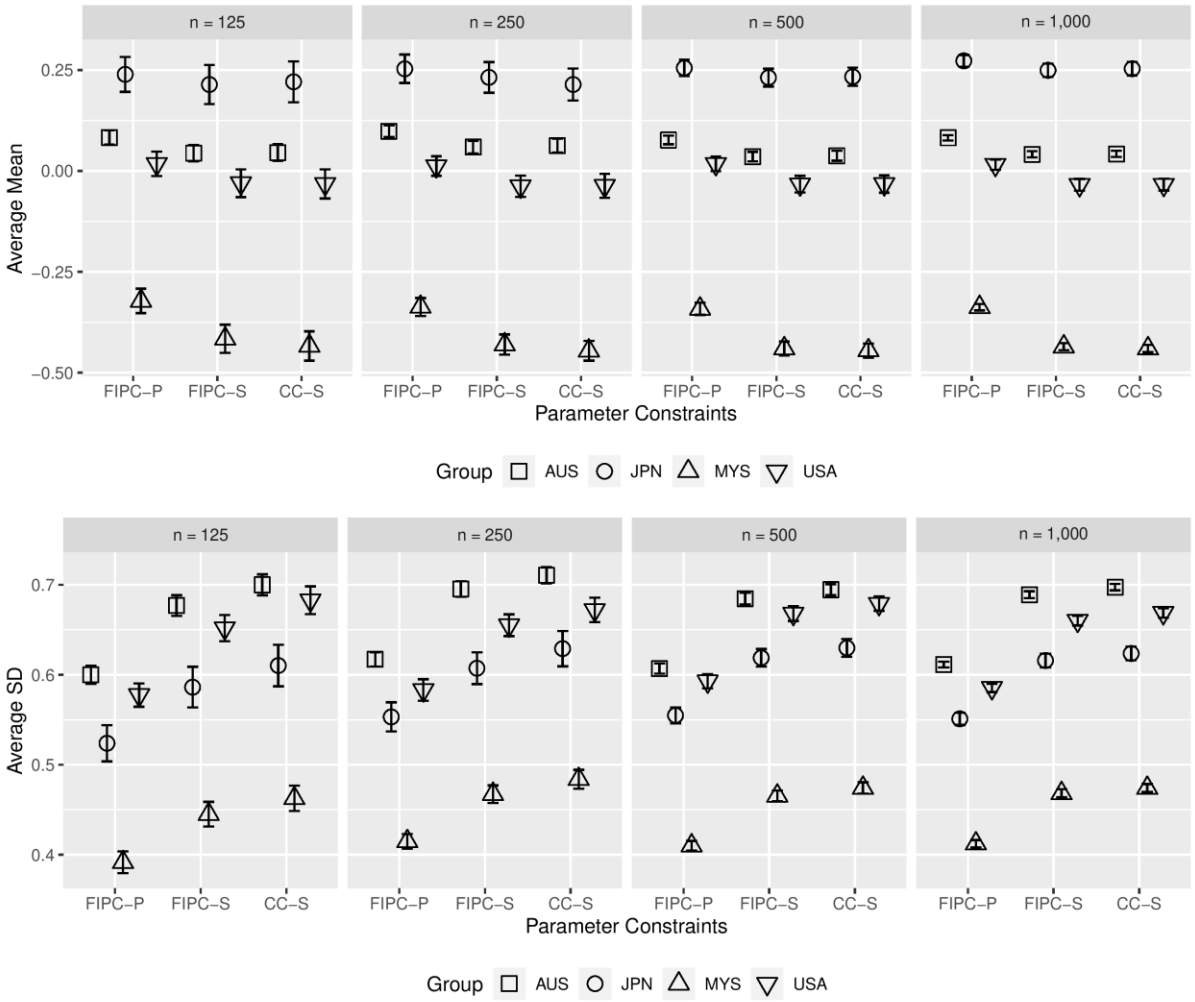

Figure 4. Differences in means and standard deviations of the country-specific latent trait distributions between types of parameter constraints (i.e., prior information introduced into the calibration) across subsample sizes, averaged across replications. Note: Ms and SDs of normal densities. Error bars represent $\pm 2SE$.

Sample size does not play a vital role for the mean of the country-specific latent trait distributions. Moreover, the marginal differences in the standard deviations of the country-specific latent trait distributions across sample sizes do not warrant substantially different conclusions about the performance of either FIPC-P or FIPC-S compared to CC-S.

Discussion and Conclusion

The purpose of this study was to investigate the impact of FIPC on the accuracy of the parameter estimates of new items in small sample situations using real data from the PISA

2015 Science assessment. The aim was to provide an indication if and how FIPC contributes to PISA's flexibility, that is, how much of the current field-trial sample size is critical for the accuracy of initial item calibration and scaling in the field trial. Therefore, we compared FIPC with different parameter constraints (i.e., different amounts of prior information introduced into the calibration) to a scaling of the study sample (CC-S) in which both trend and new Science item parameters were freely estimated and no prior information was introduced. We examined FIPC with two different constraints on the item parameter estimation. First, we fixed the item parameter estimates of trend items to their values obtained from the PISA 2015 scaling (FIPC-P). Second, we fixed the item parameter estimates of trend items to their values obtained from CC-S. The performance of FIPC was assessed by examining the bias and precision of the parameter estimates for the new Science items (machine- and human-coded), and the fit indices of the trend and new items. Moreover, the means and standard deviations of the country-specific ability distributions were compared across types of parameter constraints and subsample sizes.

Overall, FIPC-P performs better than CC-S in terms of bias of the item discrimination estimates. Parameter recovery is acceptable even with the smallest sample size. Moreover, item fit indices of trend and new Science items remain acceptable until $n = 250$. There are no differences between parameter constraints in terms of standard errors; the statistical uncertainty of the item parameter estimates increases quickly. This is a direct effect of the smaller country-specific effective sample size resulting from the complex booklet design (Mazzeo & von Davier, 2014). Overall, results show an acceptable performance of both FIPC-P and FIPC-S in country-specific samples as small as $n = 250$. Except for the standard errors of the parameter estimates of the new Science items, performance of FIPC depends primarily on the prior information introduced by the parameter constraints. This also applies to the performance of FIPC in terms of the country-specific ability distributions. The difference between FIPC-P and

CC-S can be explained by the mean discrimination and difficulty estimates of the trend Science items obtained in the PISA 2015 assessment. They are slightly increased, and because the trend items set the scale for the new Science items, their discriminations and difficulties are increased as well. This results in better performance of countries with low initial mean abilities compared to FIPC-S and CC-S.

Overall, the performance indicates that FIPC is applicable to assessments with many heterogeneous groups and a complex test design. These results complement Kim and Kolen (2019), who showed that FIPC is applicable to situations with two groups and simpler test designs.

The performance of FIPC-P can be explained as follows. First, the number of trend items is large. Arai and Mayekawa (2011) and Kim et al. (2018) show that bias in item parameter estimates decreases with an increasing number of items with parameter estimates available from previous calibrations. Second, the software used for scaling the PISA data, *mdltm* (von Davier, 2005), combines multiple updates of the prior latent ability distribution with multiple EM cycles. According to Kim (2006), multiple prior weights updating and multiple EM cycles (MWU-MEM) is the recommended configuration of the MML algorithm for FIPC. This configuration updates the latent ability distributions and the new item parameter estimates continuously until the EM algorithm converges. In the first EM cycle, however, only the estimates of the trend Science items are used to estimates the latent ability distributions. This sets the base scale for both trend and new Science items. Once the base scale is set, the following (multiple) EM cycles use both trend and new Science items to estimate the latent ability distributions (Kim, 2019). Third, using FIPC decreases the number of parameters in the estimation of the IRT models, thus decreasing the model complexity as well. Fourth, FIPC in PISA might also work well because the new and trend items can be placed on a common scale without the need for a separate equating step. Thus, large shifts in the latent ability distributions

of the content coverage of the constructs are not to be expected. While new items are developed to represent changes in the PISA frameworks, they do not measure an entirely different scale, but rather new aspects of the same scale (e.g., Science). It should be noted that before new and trend items are placed on the same scale in the PISA main survey, the dimensionality of both item groups is examined in the field trial and results are cross-validated in the main survey. In PISA 2015, it could be shown that new and trend items can be sufficiently described by a common or unidimensional IRT scale (OECD, 2017c).

In summary, there are two main benefits of FIPC over CC-S. First, FIPC directly includes linking to the base scale (e.g., the PISA trend scale) by fixing the trend item parameters to their previously estimated values. Therefore, FIPC does not require a separate equating step. Second, FIPC reduces model complexity (depending on the number of trend items, the number of parameters to be estimated decreases considerably). As shown in this study, these benefits contribute to a considerable reduction in the required sample size for an accurate scaling of large-scale assessment data.

This benefits the field trial. The field trial is critical for PISA and other international and national large-scale assessments as it is the preparation for the main survey (e.g., based on the initial field-trial item parameter estimation, items are selected or excluded for the main survey design, and different design innovations are examined). The aforementioned critical sample size of $n = 250$ students per country implies that approximately 12% of the total sample size for each country of PISA's field trial is required for the intended accuracy of the initial IRT scaling of the assessment data. It should be noted that we are not recommending to reduce the sample size of the field trial, but to use the existing sample more efficiently. With FIPC, a large part of the field-trial sample could be used to test new design features for their possible implementation in the PISA main survey without the need to increase the target sample size of the field trial while still allowing for accurate item parameter estimates. In this regard,

however, it is also important to assure the sample representativeness for each subsample. That is, the subsamples used to test different new design features and the subsample used for the preliminary field-trial item parameter estimation need to be representative of a country's population.

Although the utility of FIPC was investigated in the context of PISA, the question of how to make most use of the field trial without increasing sample size and costs is important for other large-scale assessments as well. Assessments such as Trends in International Mathematics and Science Study, Progress in International Reading Literacy Study, and the National Assessment of Educational Progress (US) also use field trials to test innovations with regard to framework, items, test design, and reported scale. Moreover, the IRT models used in these large-scale assessments are similar to, or are the same as in, PISA. Hence, the core results of this study generalize beyond PISA to other large-scale assessments as well.

In conclusion, this study illustrates how to use FIPC in order to maximize the utility of the field trial for the flexibility of the main survey. On the one hand, the findings suggest that a larger number of tests of potential new features could be examined in the field trial in preparation for the main survey without necessarily increasing the sample size. On the other, they tell us to what extent to increase the total sample size of the field trial (if necessary) to include even more features to be studied in the field trial. These are necessary requirements to keep the main survey up to date with current developments in the field of psychometrics and to meet stakeholders' needs while, at the same time, keeping costs reasonable.

References

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

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

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

Buchholz, J., & Hartig, J. (2019). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement, 43*, 241–250. https://doi.org/10.1177/0146621617748323

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Ertl, H, (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education, 32*, 619–634. https://www.jstor.org/stable/4618685

Fischman, G. E., Marcetti Topper, A., Silova, I., Goebel, J., & Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national educational policies. *Journal of Education Policy, 34*, 470–499. https://doi.org/10.1080/02680939.2018.1460493

Frey, A., & Hartig, J. (2020). Methodological challenges of international student assessment. In H. Harju-Luukkainen, N. McElvany, & J. Stang (Eds.), *Monitoring of student achievement in the 21st century—European policy perspectives and assessment strategies*. Springer.

Frey, A., Hartig, J., & Rupp, 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*, 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Gonzalez, E. J., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In M. von Davier, & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 3) (pp. 125–156). IEA-ETS Research Institute.

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. https://doi.org/10.1177/0146621602026001001

Hopfenbeck, T., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*, 333–353. https://doi.org/10.1080/00313831.2016.1258726

Kang, T., & Petersen, N. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review, 12*, 311–321. https://doi.org/10.1007/s12564-011-9197-2

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. https://doi.org/10.1177/0013164410375111

Keller, L. A., & Keller, R. R. (2015). The effect of changing content on IRT scaling methods. *Applied Measurement in Education, 28*, 99–114. https://doi.org/10.1080/08957347.2014.1002922

Keller, R. R., Keller, L. A., & Baldwin, S. (2007). *The effect of changing equating methods on monitoring growth in mixed-format tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software mdltm including parallel EM algorithm. In M. von Davier & Y-S. Lee (Eds.), *Handbook diagnostic classification models* (pp. 603–628). Springer.

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

Kim, S., Cole, K. L., & Mwavita, M. (2018). FIPC linking across multidimensional test forms: Effects of confounding difficulty within dimensions. *International Journal of Testing, 18*, 323–345. https://doi.org/10.1080/15305058.2018.1428980

Kim, S., & Kolen, M. J. (2019). Application of IRT fixed parameter calibration to multiple-group test data. *Applied Measurement in Education, 32*, 310–324. https://doi.org/10.1080/08957347.2019.1660344

Kim, Y. (2019). Two IRT fixed parameter calibration methods for the bifactor model. *Journal of Educational Measurement, 57*, 29–50. https://doi.org/10.1111/jedm.12230

Lenth, R. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.4. https://CRAN.R-project.org/package=emmeans

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mazzeo, J., & von Davier, M. (2014). Linking scales 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. 229–1258). CRC Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–177. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Organisation for Economic Co-operation and Development (2017a). Test design and development. In OECD (Ed.), *PISA 2015 technical report* (pp. 30–55). OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA-2015-Technical-Report-Chapter-2-Test-Design-and-Development.pdf

Organisation for Economic Co-operation and Development (2017b). Sample design. In OECD (Ed.), *PISA 2015 technical report* (pp. 66–89). OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA-2015-Technical-Report-Chapter-4-Sample-Design.pdf

Organisation for Economic Co-operation and Development (2017c). Scaling PISA data. In OECD (Ed.), *PISA 2015 technical report* (pp. 128–185). OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/09_Chapter_09_PISA2015.pdf

Organisation for Economic Co-operation and Development (2017d). International Data Products. In OECD (Ed.), *PISA 2015 technical report* (pp. 375–382). OECD Publishing. http://www.oecd.org/pisa/data/2015-technical-report/19_Chapter_19_PISA2015.pdf

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. https://doi.org/10.1207/s15324818ame1802_4

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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

Rutkowski, D. (2018). Improving international assessment through evaluation. *Assessment in education: Principles, Policy & Practice, 25*, 127–136. https://doi.org/10.1080/0969594X.2017.1300572

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

Von Davier, M. (2005). Multidimensional discrete latent trait models (mdltm) [Computer software]. Educational Testing Service.

Von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28.

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

Waldow, F. (2009). What PISA did and did not do: Germany after the 'PISA-shock.' *European Educational Research Journal, 8*, 476–483. https://doi.org/10.2304/eerj.2009.8.3.476

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2018). Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling, 60*, 145–164.

Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling, 60*, 347–368.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology, 8*, 484–495. https://doi.org/10.3389/fpsyg.2017.00484