Organisation for Economic Co-operation and Development

EDU/WKP(2022)15

Unclassified **English text only**

17 August 2022

DIRECTORATE FOR EDUCATION AND SKILLS

The impact of growing participation in PISA on scaling outcomes. A Monte Carlo simulation study.

OECD Education Working Paper No. 277

Artur Pokropek, Marek Muszyński, and Tomasz Żółtak (Institute of Philosophy and Sociology, Polish Academy of Sciences)

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Artur Pokropek, artur.pokropek@gmail.com

JT03501196

OECD EDUCATION WORKING PAPERS SERIES

---

www.oecd.org/edu/workingpapers

Abstract

The OECD's PISA (Programme for International Student Assessment) is expanding on new educational systems from cycle to cycle. Most of the new participants differ significantly and negatively from the core participants in the level of educational proficiency. The study has investigated a potential expansion of the new participants' proportion from 15% of low-performing countries to 50% of low-performing countries in the PISA sample. A simulation study was performed that aimed to check whether an increasing share of low-performing countries among PISA participants can affect: a) key country parameters (means and within-country standard deviations), b) item parameters (difficulty and discrimination), and c) sensitivity and specificity of differential item functioning procedures. The results of the study point out that the PISA procedures are fit and robust to the increasing proportion of low-performing countries resulting in highly reliable inter-country score differences and estimates of country means.

Keywords: PISA, international large-scale assessments, low-performing countries in PISA, differential item functioning (DIF), country and item parameter recovery, heterogeneous populations

Executive summary

The OECD's PISA (Programme for International Student Assessment) is expanding on new educational systems from cycle to cycle; from the first edition in 2000 to the forthcoming assessment in 2022 the number of participating entities doubled (from 43 to 87). Most of the new participants differ significantly from the core participants in the level of educational proficiency. Typically, countries that joined PISA recently are characterised by low educational attainment. The expectation is that 10 to 20 new educational systems will join PISA every cycle. It means that low-performing countries will soon form a large part of PISA participants.

To scale item, student, and country data PISA employs a state-of-the-art set of psychometric tools based on item response theory. As it is always the case in statistics, these tools bring a group of assumptions and technical restrictions, violations of which may result in biased estimation of key parameters. The increasing proportion of low-performing countries among PISA participants inflates their heterogeneity and raises concerns about the assumptions of normal proficiency distribution and high student-test match that would ensure precise estimation of model parameters. Thus, the adequacy of the currently used PISA scaling procedures is questioned as they are now used in a more and more diversified set of participants. The following study was planned to check if the quality of the PISA-based indicators is in any way threatened by an ever-expanding roster of PISA participants.

This question was addressed in a simulation study that aimed to check whether an increasing share of low-performing countries among PISA participants can affect: a) key country parameters (means and within-country standard deviations), b) item parameters (difficulty and discrimination), and c) sensitivity and specificity of differential item functioning procedures in Item Response Theory (IRT) modelling of PISA data. The study comprised several simulation conditions. Most importantly, the proportions of 15%, 35%, and 50% of low-performing countries among PISA participants were compared with each other regarding the key study indicators. It was also checked whether bias in item parameters was related to their difficulty. Precise estimation of parameters of very easy or very difficult items may be difficult, because of low variance in responses to such items.

Because the study aimed at assessing properties of the PISA scaling model and not the population model used to compute final plausible values (PVs) estimates of individuals' proficiency, we ruled on using data from only one domain (science) to reduce estimation complexity and time. Science was selected as it was the main domain in the PISA 2015 cycle, the cycle on whose design the simulation was based.

Additionally, some simple remedies for the previously observed problems related to scaling low-performing countries data were tested. These remedies included introducing easier items to PISA, scaling data from high-performing participants first, and improving the existing differential item functioning detection procedure. The first condition included comparison between the country and item parameters estimated from the item pool of normal PISA difficulty and the item pool with the difficulty of non-linking items lowered by 100 PISA points. The second consisted of comparing effects of estimating the item parameters basing solely on the data from high-performing countries with item calibration where data from both types of participants are used simultaneously. The third condition comprised analysing possible gains of having a hypothetical perfect procedure of identifying differential item functioning in comparison to the currently used procedures.

The results of the whole study point out that the PISA procedures are fit and robust to the current proportion of low-performing countries, but also to a proportion as high as 50%

of low-performing countries in the PISA sample. This means that highly reliable country score differences and estimates of country means were achieved in all tested conditions. Therefore, it can be justifiably claimed that PISA expansion to the group of low-performing countries does not pose any serious threat to the precision of the key PISA indicators: country means and within-country standard deviations, and neither to inter-country score differences.

Although inter-country score differences of proficiency were recovered very well, we observed that estimates of country mean proficiency and standard deviations obtained from the IRT models are overestimated. The effect is systematic and therefore does not affect the ordering of the groups. What is most important from the perspective of this study is that the size of this effect does not depend on the share of low-performing countries in the data. Moreover, it is important to notice that the effect would be most probably mitigated by a complete application of operational PISA scaling procedures, most importantly generating plausible values (PVs) from all three domains (mathematics, reading, science) together.

Despite the large robustness of PISA IRT scaling methodology to the increasing number of LPCs confirmed by this study, the PISA scaling procedures can be further improved, most notably by investigating differential item functioning more deeply and testing new ways of its handling in scaling procedures. Specifically, it is crucial not only to detect DIF but also to use good criteria whether available data enables estimating group-specific parameters of an item with a reasonable precision. If there are too few respondents and/or too low variation in responses, trying to estimate item parameters freely may cause more harm than ignoring DIF in the model specification. Developing procedures for better matching test difficulty and assessed groups abilities should also be a priority. Such procedures may entail adaptive testing that was already implemented in PISA and developing a large item bank with sufficient number of items for all ranges of difficulty. This study did not investigate population modelling and generating PVs which arguably should be the direction for further methodological research on PISA.

List of abbreviations

CBA - computer-based assessment

DIF - differential item functioning

ILSAs - international large-scale assessments

LPCs - low-performing countries

MAE - mean absolute error

OECD - Organisation for Economic Co-operation and Development

PBA - paper-based assessment

PISA - Programme for International Student Assessment

PVs - plausible values

RMSE - root-mean squared error

Table of contents

Abstract ... **3**

Executive summary ... **4**

List of abbreviations ... **5**

1. Introduction ... **8**

2. Simulation conditions ... **10**

2.1. The proportion of low-performing countries (LPC) ... 10
2.2. Difficulty of the test ... 11
2.3. Test scaling ... 11
2.4. Detection and treatment of country-by-item interactions (DIF) ... 12
2.5. Summary of the conditions ... 12

3. Results ... **13**

3.1. Impact of increasing number of low-performing countries on PISA scaling ... 14
3.1.1. Inter-country score differences and group parameters ... 14
3.1.2. Item discrimination and difficulty recovery ... 16
3.1.3. DIF detection ... 19
3.2. Can we do better? ... 20
3.2.1. Difficulty of the test ... 21
3.2.2. Test scaling ... 22
3.2.3. Perfect DIF detection ... 24
3.2.4. Ignoring DIF ... 25

4. Discussion ... **27**

5. Method in details ... **28**

5.1. Data generation procedures ... 28
5.1.1. Group characteristics ... 29
5.1.2. Item characteristics ... 31
5.1.3. Data simulation ... 32
5.2. Estimation procedures ... 33

References ... **35**

Annex A. Additional graphs for simulation results and simulation code ... **39**

Additional graphs for 3.1 ... 39
Additional graphs for 3.2 ... 42
Simulation code ... 45

Tables

Table 1. Pearson's correlation between true country means and estimated country means 14
Table 2. Pearson's correlation between true country means and estimated country means in conditions with lower item difficulties of non-linking items. Reference values, for current type of scaling, in parentheses 21
Table 3. Pearson's correlation between true country means and simulated country means in conditions with first calibration performed using only OECD countries. Reference values, for current type of scaling, in parentheses 23
Table 4. Pearson's correlation between true country means and estimated country means with perfect DIF detection. Reference values, for current type of scaling, in parentheses 24
Table 5. Pearson's correlation between true country means and estimated country means in conditions with ignoring DIF. Reference values, for current type of scaling, in parentheses 26
Table 6. Countries used as donors for the simulation 30

Figures

Figure 1. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants 15
Figure 2. Bias and RMSE for country means as a function of different proportion of LPCs and countries' proficiency levels 16
Figure 3. Recovery of item parameters for different proportion of LPCs 17
Figure 4. Bias of difficulty parameter and differences between item difficulty (in a given country in the data generating model) and country mean proficiency (in the data generating model) 18
Figure 5. Bias of discrimination parameter and differences between item difficulty (in a given country in the data generating model) and country mean proficiency (in the data generating model) 19
Figure 6. Sensitivity and Specificity of DIF detection for different proportion of LPCs 19
Figure 7. Sensitivity and Specificity of DIF detection for different proportion of LPCs and countries' mean proficiency 20
Figure 8. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with lower item difficulties of non-linking items 22
Figure 9. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with first calibration performed using only OECD countries 23
Figure 10. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with Perfect DIF detection 25
Figure 11. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with ignoring DIF (calibrating only the invariant model) 27
Figure 12. Structure of the simulation design 29

Figure A A.1. MAE for country means for different proportion of LPCs and countries' mean proficiency 39
Figure A A.2. Bias for country standard deviations for different proportion of LPCs and countries' mean proficiency 40
Figure A A.3. MAE for country standard deviations for different proportion of LPCs and countries' mean proficiency 40
Figure A A.4. RMSE for country standard deviations for different proportion of LPCs and countries' mean proficiency 41
Figure A A.5. Recovery of item parameters for different proportion of LPCs in conditions with lower item difficulties of non-linking items 42
Figure A A.6. Recovery of item parameters for different proportion of LPCs in conditions with first calibration performed using only OECD countries 42
Figure A A.7. Recovery of item parameters for different proportion of LPCs in conditions with Perfect DIF detection 43
Figure A A.8. Recovery of item parameters for different proportion of LPCs in conditions with ignoring DIF (calibrating only the invariant model) 43
Figure A A.9. Recovery of DIF for different proportion of LPCs in conditions with lower item difficulties of non-linking items 44
Figure A A.10. Recovery of DIF for different proportion of LPCs in conditions with first calibration performed using only OECD countries 44

1. Introduction

Due to the increasing demand for evidence-based policy in education, more and more countries join international large-scale assessments (ILSAs), such as PISA. As indicated by Kamens and McNeely (2010[1]), more than 50%[1] of the world's countries have already participated in some kind of ILSA and the number of participants has further increased since then. With every PISA cycle, 10-20 new participants are expected, most of them participating in an international assessment for the first time. This influx to ILSAs results in PISA participants being more and more diverse, not only in terms of economic or cultural factors, but also regarding their educational proficiency. This affects PISA, once aimed at a small group of wealthy, developed countries, and now being administered to an increasingly heterogeneous sample of participants, with growing discrepancy in educational achievement among the assessed countries (Lockheed, Prokic-Bruer and Shadrova, 2015[2]; Rutkowski and Rutkowski, 2018[3]; Rutkowski, Rutkowski and Liaw, 2018[4]; Rutkowski, Rutkowski and Liaw, 2019[5]).

This rising heterogeneity brings forth several technical issues. First of all, there is a problem of measuring the low-performing countries' (LPCs) achievement adequately. To tackle this problem, clusters of easier items were introduced to PISA booklets in order to assess LPCs' proficiency more accurately. Moreover, an adaptive test design in the form of multistage testing was adopted, to ensure more precise measurement across a wide range of student performance. However, it seems that the easier items were still too few and too difficult (Rutkowski, Rutkowski and Liaw, 2018[4]; Rutkowski and Rutkowski, 2021[6]; Rutkowski, Rutkowski and Liaw, 2019[5]) to provide more precise measurement for the LPC students' proficiency. What is more, the currently used procedure to identify differential item functioning (DIF) based on the RMSD (root-mean-square deviation) statistic (Khorramdel, Shin and von Davier, 2019[7]) proved to be suboptimal for detecting DIF in LPCs, leading to concerns of possible achievement underestimation in this group of countries (Tijmstra et al., 2020[8]).

Apart from these concerns, there is also a more general concern regarding how an increasing proportion of LPCs in PISA will affect the measurement quality of PISA-based indicators. Further PISA expansion can put into question operational procedures currently used in the PISA data scaling. A large difference (mismatch) between respondents' proficiency and item difficulty could increase measurement errors and reduce estimation accuracy (Rutkowski, Rutkowski and Liaw, 2018[4]). Moreover, model assumptions could be violated in terms of item (Guenole and Brown, 2014[9]; Köhler and Hartig, 2017[10]) and model (Oliveri and Von Davier, 2011[11]; Sinharay and Haberman, 2014[12]) misfit. However, previous studies have not investigated the impact of these problems on key PISA results. They focused on DIF detection – e.g. (Tijmstra et al., 2020[8]) – or accurate recovery of the ability estimates in the group of low-performing countries (LPCs) – e.g. (Rutkowski, Rutkowski and Liaw, 2018[4]) – but not on a potential impact on PISA country ranking, group characteristics or effects of including more LPCs in PISA on the estimates of the rest of the participating countries.

Taking all of the above into consideration, we propose a simulation study that aims at answering the question: if, and how, does the rising proportion of LPCs among the PISA participants affect key PISA indicators: country means and standard deviations

[1] See more on educational data availability among the world's countries: http://uis.unesco.org/sites/default/files/documents/sdg4-data-digest-2019-en_0.pdf (Chapter 1) and http://uis.unesco.org/sites/default/files/documents/sdg4-databook-global-ed-indicators-2019-en.pdf (Figure 1)

(populational distributions) and, consequently, the inter-country score differences[2]. We will also investigate how the increased share of LPCs affects item parameters recovery (item difficulty and discrimination) and how it is related to the capabilities of DIF detection procedures that is currently used in PISA. Additionally, we show the results as a function of the country proficiency level to better understand the potential impact of the increasing number of LPCs on key PISA parameters. The quality of item parameter recovery is also presented as a function of item difficulty. The results of our analysis are most often displayed separately for LPCs and high-performing countries referred to as "OECD countries" in the remainder of this study[3].

Finally, we consider modifications for the current methodology, investigating how the quality of the PISA results could be maintained in the presence of an increasing number of LPCs. First, we check how substantial decrease of the difficulty of the non-linking items[4] would affect the overall quality of the PISA scaling. Second, we test whether concurrent *versus* separate calibration of OECD and LPC countries would be more effective. Third, by introducing a condition with perfect DIF detection (i.e. assuming that it is known which items in which countries are affected by DIF), we checked how much could be gained by improving current PISA DIF detection procedures (which for now do not promise perfect DIF detection).

This study is focused on the IRT scaling model used in PISA. We are not investigating population modelling that involves conditioning on background variables, using item parameters from all domains concurrently and generating PVs (Mislevy, 1991[13]; Mislevy and Sheeham, 1987[14]; Thomas, 2002[15]; von Davier et al., 2006[16]; Von Davier, Gonzalez and Mislevy, 2009[17]). This additional step was designed to improve the accuracy of the group level estimates in different ILSAs (as well to simplify the secondary analysis). Therefore, the results presented here could be viewed as a worst case scenario situation as using all three domains would almost triple the number of items used in estimations and, consequently, greatly reduce the sheer size of the observed biases (Thomas, 2002[15]). Using conditioned PVs to calculate country means would also result in larger similarity between our results and the operational PISA (Zieger et al., 2020[18]).

[2] By inter-country score differences we mean the difference between the countries' means from the generated data and the countries' means from the estimated simulation condition. A high correlation between the two indicates a high precision of estimates and suggests good recovery of the country rankings.

[3] We refer to the high-performing countries as "OECD countries" throughout the report for the sake of clarity and brevity. Of course, we acknowledge that among the OECD members there are also countries with a mean performance level close to the 400 PISA points threshold of an LPC. We also use the term "country" to denote all kinds of entities participating in PISA for the same reasons of brevity.

[4] The item difficulty was lowered for all participants. In this way this analysis is different to the ones described; e.g. (Rutkowski and Rutkowski, 2021[6]) and similar studies.

2. Simulation conditions

To generate the data for simulations we started with original PISA 2015 data for the science domain to exactly mimic the structure of the assessment. Therefore, group characteristics used in the simulation imitate the real data situation as far as it is possible. Item parameters used for the simulation were also taken from real data; some of them were kept as estimated in PISA 2015 study, some were changed, depending on simulation conditions.

We modelled the full test design, sample design and respondents' missing data patterns as they are in the real PISA implementation. The countries from which we obtain information are referred to as donor countries. The simulated groups reflecting properties of the real countries are called recipients. Donor countries were selected from PISA 2015 data. To reduce complexity, we selected countries where the majority of students used one testing language (at least 70% of participants sitting for an assessment in the dominant language). Only data and characteristics of this dominant language group were used for each country. Countries were divided into two clusters – non-LPC (high-performing countries, referred to in the report as OECD countries) and LPC countries (countries with an average PISA science score below 400 points). Please note, that this procedure results in means of simulated countries different from the original means of donor countries and that the simulated means differ between the conditions by design. This is because country means were adjusted in each of the simulation conditions in order to achieve a twofold goal: a) to obtain the assumed number of LPCs (countries with mean proficiency below 400 PISA points), b) to keep the average proficiency of non-LPCs constant across conditions. Detailed description of data generating procedures are presented in point 5 of the report.

2.1. The proportion of low-performing countries (LPC)

LPCs are defined as a non-OECD PISA participant scoring below 400 PISA points. Based on the results achieved by countries that have recently (2015 and 2018 cycles) joined PISA and based on the results of the PISA-D participants (Rutkowski and Rutkowski, 2021[6]), we assumed that this is the most probable characteristic of countries joining PISA in next cycles. We assume that the proportion of low-performing countries to all participating countries is a key characteristic affecting main study results. The results are simulated for three proportions, selected to show the current state-of-the-art (PISA 2018), the state expected in the near future (PISA 2025), and a probable state somewhere beyond the year 2030. The number of entities included in simulation study, 26, is similar or larger to the numbers used in similar simulation studies (Rutkowski, Rutkowski and Zhou, 2016[19]).

a. proportion **15%** - similar to PISA 2018 (22 OECDs + 4 LPCs)
b. proportion **35%** - as expected in PISA 2025 (17 OECDs + 9 LPCs)
c. proportion **50%** - as expected beyond PISA 2030 (13 OECDs + 13 LPCs)

It is to be remembered that up to a half of the LPCs are countries that participated in the paper-based assessment (PBA), hence they contribute only to some item parameter estimates and whose group characteristics are based on only a subset of items in comparison to the countries that participated in the computer-based assessment (CBA). This is because PBA contained a lower number of items than the CBA. We see these conditions as relevant, because at least some LPCs joining PISA in the future may prefer the PBA given technical and organisational constraints.

We do not plan to investigate mode effects directly as this is not the aim of the study. The general consensus is that the mode effects in PISA are of minor role (OECD, 2017[20];

Von Davier, Forthcoming[21]) although with some concerns (Robitzsch et al., 2020[22]). It seems interesting to test mode effects under varying proportions of low-performing countries in future simulation studies.

2.2. Difficulty of the test

Two conditions were considered for the difficulty of the test**:**

a. Normal item difficulty (testing as in PISA 2015);

b. Assuming that the non-linking item pool will be easier (by 100 points on the PISA scale) than in the current PISA (keeping linking items' difficulty unchanged).

In the **b.** condition we assume that difficulty of non-linking items in the PISA's item pool is lowered by 100 PISA points. In this condition, the linking items' difficulty remains unchanged, hence the overlap between the pools - the real PISA item pool and the simulated easier item pool - is large. This change should allow better modelling of the LPCs' performance, as their average performance is now significantly below the difficulty of the PISA item pool (Rutkowski and Rutkowski, 2021[6]). However, this change affects only countries in which computer-based assessment was used, because the paper version of the PISA 2015 test included linking items only. Easier items were in fact introduced in PISA 2009 and 2012 cycles for reading and mathematics (one easier cluster per student), but its effect on measurement precision was never evaluated (Rutkowski, Rutkowski and Liaw, 2018[4]). Additionally, easier items were available also in PISA 2015, but only for minor domains of this cycle (mathematics and reading). Importantly, the easier items were delivered only in countries that volunteered for the easier booklets. However, even in these countries only a limited number of students were actually administered easier items (around 18% of students whose tests had 25% items were easier items and about 3% of students whose tests had 50% easier items (Rutkowski and Rutkowski, 2021[6]); in effect most of the students still sit for the test of standard difficulty. It was evidenced that introducing easier items would help to assess LPCs proficiency (Rutkowski, Rutkowski and Liaw, 2018[4]; 2019[5]); here we aim to evaluate how it affects other countries' parameters.

2.3. Test scaling

There are widespread concerns that DIF occurs more frequently in LPCs (Rutkowski and Rutkowski, 2021[6]; Tijmstra et al., 2020[8]) and, consequently, including more LPCs may result in biased initial estimates of item parameters, obtained from a model with invariant specification. As a remedy, one may first estimate item parameters using only OECD countries data, so the initial estimates were not affected by DIF in LPCs. Only then one would search for DIFs in all the countries using the current PISA DIF detection procedure. This approach was in fact used in PISA up to the 2012 edition – see p. 155 of (OECD, 2014[23]). To assess the impact of this scenario, we compared two conditions:

a. using response vectors from all countries in the initial item parameter estimation

b. using response vectors from the OECD countries only in the initial item parameters calibration and including data from LPCs only afterwards in the DIF detection procedure (similarly to the procedure used for paper-based assessment (PBA) in PISA 2015).

2.4. Detection and treatment of country-by-item interactions (DIF)

Three DIF detection procedures were compared in this simulation condition:

a. Assuming perfect DIF detection

b. Assuming current DIF detection procedures (using RMSD)

c. Ignoring DIF (i.e., using estimates from the invariant model only).

In the **a.** condition, we assume that we know which items in which countries are affected by DIF. Results obtained in this condition depict the best situation possible and show how much we could possibly gain in proficiency estimates accuracy by developing more accurate DIF detection procedures. In the **b.** condition, we carefully model the procedure currently used to detect DIF items in PISA – as in PISA 2015 technical report and personal information obtained from researchers performing these analyses (OECD, 2017[20]). Finally, the **c.** condition will enable us to show how biased the results would be if we were to ignore DIF modelling entirely, showing the benchmark for one of the most unfavourable situations from the perspective of model misspecification.

2.5. Summary of the conditions

In total, there are **six simulation conditions (3 proportions of LPCs times 2 test difficulty distributions) that refer to the data generating process**. For each of the conditions the number of estimated models is determined as follows:

a. one model assuming perfect DIF detection (in this case there is no point in scaling responses from OECD and LPC countries separately)

b. one starting model assuming current DIF detection procedure and using the response vectors from all countries in item parameter estimation (results from this starting calibration are used in the analysis as the "ignoring DIF" condition), followed by several re-estimations discarding the assumption of a fixed value for some parameters (freeing parameters) in some groups using the RMSD criterion

c. two starting models assuming current DIF detection procedure and using the response vectors from the OECD countries and the low-performing countries in separate parameter calibrations, followed by several re-estimations freeing parameters in some groups using the RMSD criterion.

Assuming that at least three re-calibrations will be needed for each of the two variants of the test scaling procedure we have to estimate a minimum of 6*(1+ (1+3) + (2+3)) = 60 sets of model parameters for each of the 200 replications.

3. Results

We put a primary emphasis on PISA country means, investigating whether including a large number of LPCs could lower the quality of the recovery of country means and inter-country mean differences. To investigate this aspect, we inspect Pearson's correlation of the country means that were used to generate data with the country means estimated in this study. The correlation is presented for the changing proportion of LPCs among PISA participants. According to recommendations by Muthén and Asparouhov (2014[24]; 2018[25]), a correlation between the true means and their estimates of at least 0.99 indicates a reasonably good recovery of PISA rankings and, consequently, low bias in estimating each group's (e.g. country participating in a PISA cycle) latent mean. Muthén and Asparouhov (2018[25]) also showed that such correlations could be directly linked to the limit of the estimation of standard error of the estimated mean for the group[5].

To investigate the results in more detail for each model, we investigate bias and accuracy of the estimated latent group means, within-group standard deviations and item parameters. Accuracy is a combination of bias and variability that quantifies the overall performance of an estimator. The more biased and the less precise an estimator is, the worse is its accuracy. In this study, we employ two measures for assessing estimator's accuracy: the mean absolute error (MAE)[6] and the root mean square error (RMSE, see similar choice of indices in Rutkowski, Rutkowski and Liaw (2018[4]; 2019[5]). The MAE reports the average absolute differences between an estimated parameter (proficiency mean or item difficulty/discrimination) and a true parameter value (group proficiency mean or item difficulty/discrimination used to generate the data). The RMSE squares these differences (errors) instead of taking the absolute value, averages them, and then takes the square root of this average to turn back onto the scale of a given parameter. Compared to the MAE, the RMSE puts more emphasis on large differences even if these occur very infrequently.

Setting a reasonable cut-off for RMSE and MAE is not straightforward. Following previous simulation studies (Pokropek, Davidov and Schmidt, 2019[26]), we could set an "orientation point" where the cut-off should be placed. This is most easily done for the group (e.g. country) means. The standard deviation of group means used for data generation in this study was 55 PISA points. We interpret 20% of this standard deviation that is 11 PISA points, in MAE or RMSE as large average deviations.

Finally, we investigate the DIF detection by investigating sensitivity and specificity. Sensitivity in the context of DIF detection is a ratio of correctly detected DIF parameters to all DIF parameters while specificity is the ratio of correctly assessed non-DIF parameters to all non-DIF parameters. The former statistics indicates its efficacy in detecting DIF, whereas the latter indicates its efficiency (whether this is not at the expense of qualifying too many invariant parameters as being DIF). Low values of both indexes are associated with less precise estimates. Low sensitivity means that a large number of true DIF parameters are not recognised which leads to direct problems with comparability and misfit of the model. Low specificity means that many item parameters are wrongly treated as DIFs and are needlessly estimated separately for groups, which leads to the reduction of number

[5] For instance, to achieve an absolute error limit of 0.277 (in standardised metric) for 95% of the groups, a correlation of 0.99 is required, while for a slightly smaller correlation of 0.98 the limit of the estimation error is much higher and equals 0.392. On the other hand, the limit of the absolute error linked to a correlation of 0.995 is .196 and for 0.999 it equals .088 (Muthén and Asparouhov, 2018[25]):), indicating a very precise estimation.

[6] Sometimes also known as "mean absolute difference" (MAD); see for example (Rutkowski, Rutkowski and Liaw, 2018[4]).

of observations per estimated parameter and lowers the precision of estimation of a given model parameters.

3.1. Impact of increasing number of low-performing countries on PISA scaling

In this analysis we present the impact of the increasing number of low-performing countries (LPCs) in PISA on fundamental properties of the PISA rankings and indices presented in this assessment: country mean and within-country standard deviation. In this part, the only manipulated variable is the proportion of LPCs in the PISA sample. The DIF detection procedure is similar to the one used in an ordinary study (with some minor differences that are discussed later) and item difficulty remains as it is in an ordinary PISA study.

3.1.1. Inter-country score differences and group parameters

Table 1 displays Pearson's correlation between the true country means and estimated means, as well as correlations between the true and estimated within-country standard deviations. The results indicate clearly that both with the current proportion of LPCs and with an increased proportion of LPCs in the future, the recovery of rankings is very good, even slightly better in situations with a substantial proportion of LPCs in the PISA sample. Recovery of the within-country standard deviations are substantially lower – similarly to, for example, (Marsh et al., 2018[27]; Rolfe, 2021[28]) – showing that ordering a group of countries based on their within-country variation is very error-prone.

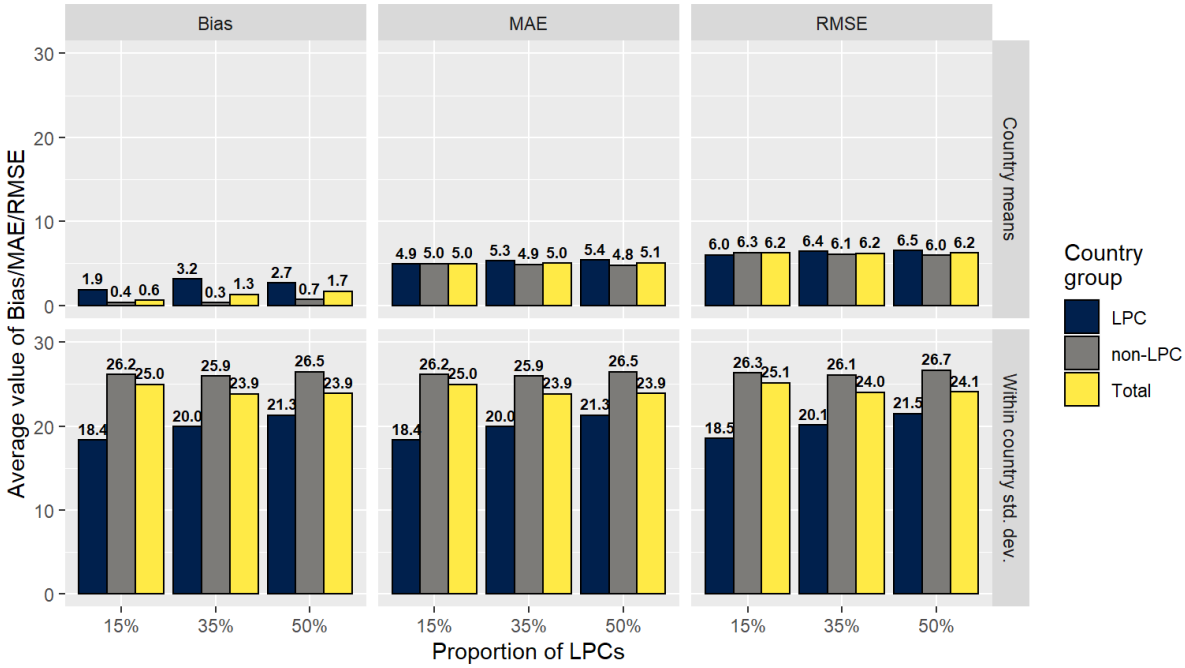
Table 1. Pearson's correlation between true country means and estimated country means

	Share of LPCs		
Group characteristics	**15%**	**30%**	**50%**
Country means	0.993	0.996	0.996
Country standard deviations	0.912	0.939	0.947

To some extent, the good recovery of the inter-country score differences is not surprising, as with an increasing proportion of LPCs in the sample, the variation between country means also increases, making these differences easier to recover. However, one should take into account that in our simulations we have also accounted for all potential problems (like DIF) in the estimation procedures. It is also for future studies to determine the consequences of having more than 50% of LPCs in the PISA sample.

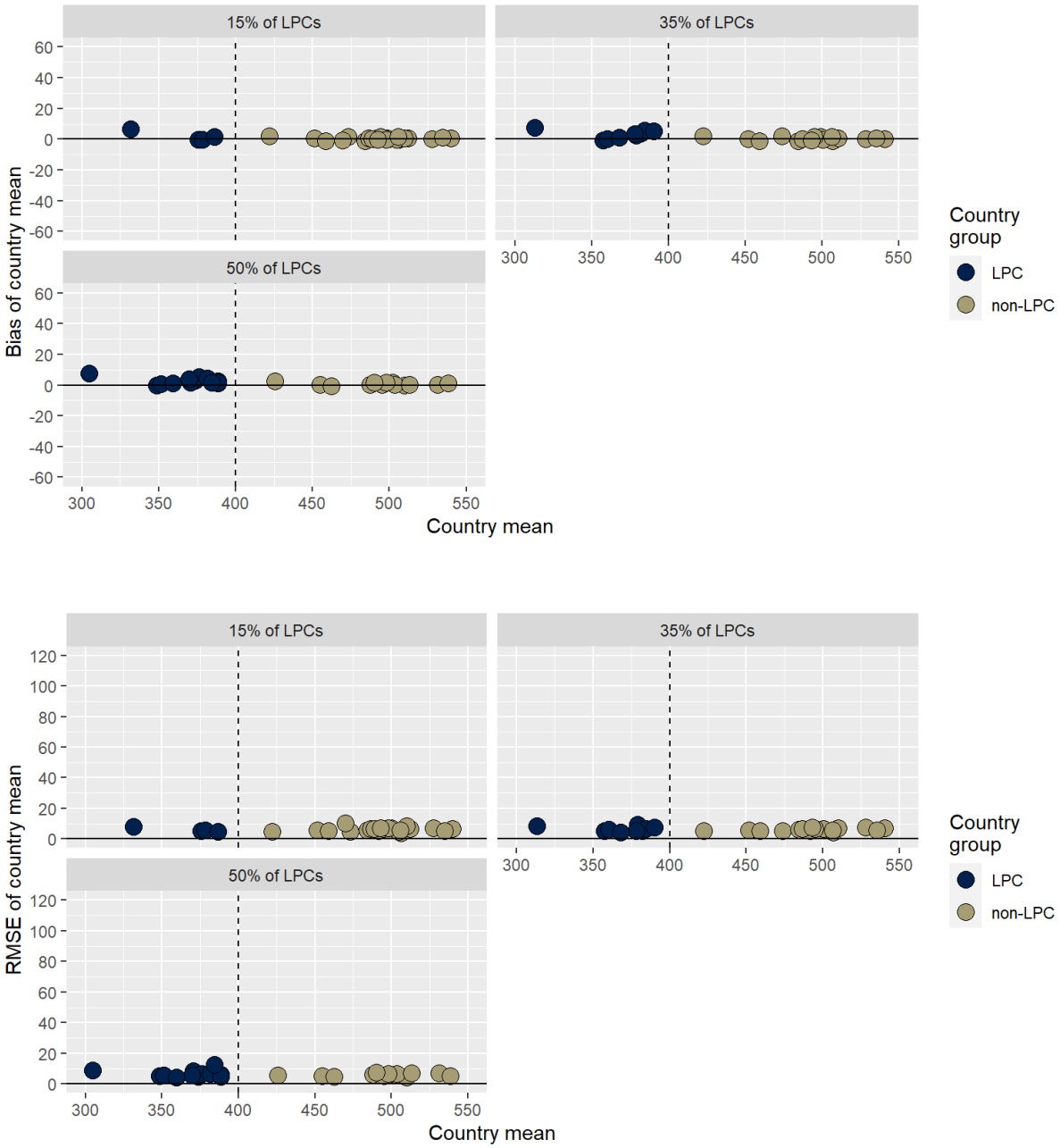
The correlation between true and estimated (simulated) country means reached the threshold of 0.99, as recommended by Muthén and Asparouhov (2014[24]; 2018[25]), while the correlations between true and estimated standard deviations proved to be lower than recommended in the literature (below 0.99). A more detailed picture could be found in Figure 1, where Bias, MAE and RMSE values are presented both for country means and within-country standard deviations, both presented for the total PISA sample (all participating countries together) and also separately for OECD and LPCs groups.

Figure 1. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants



The results indicate that country means are biased upwards, however, the size of the bias is limited to just several PISA points. MAE and RMSE values are considerably below the assumed threshold of 11 PISA points (20% of between-country variation in means). Within-country standard deviations are biased upwards in both groups of countries, with the size of the bias an order of magnitude larger as the bias in country means. This is not unusual in cross-country comparisons. For instance, in a simulation study performed by Marsh and colleagues – see Table 8 of (Marsh et al., 2018[27]) – the average mean square error (MSE) for standard deviation was 2.7 larger than the average MSE for group means. In real PISA data biases in group means and standard deviations estimations are likely to be slightly lower than those presented here. First of all, operational procedures use larger sample sizes (at least for some countries), and DIF items are clustered in groups and estimated simultaneously which gives larger sample sizes per parameter. This will however not affect the most important result that could be inferred from the Figure 1, which states that, neither the size, nor the direction of the bias seem to be related to the proportion of LPCs among PISA participants. Therefore, PISA expansion to the group of low-performing countries does not pose any serious threat to the precision of the key PISA indicators: country means and within-country standard deviations, and neither to inter-country score differences as it was demonstrated in Table 1. Of course, this conclusion only holds for proportions of LPCs to other countries as presented in this report; the impact of the share of LPCs higher than 50% of the whole sample on PISA estimations remains to be tested. Correct recovery of key country parameters is not strictly related to the country's means as presented in Figure 2.

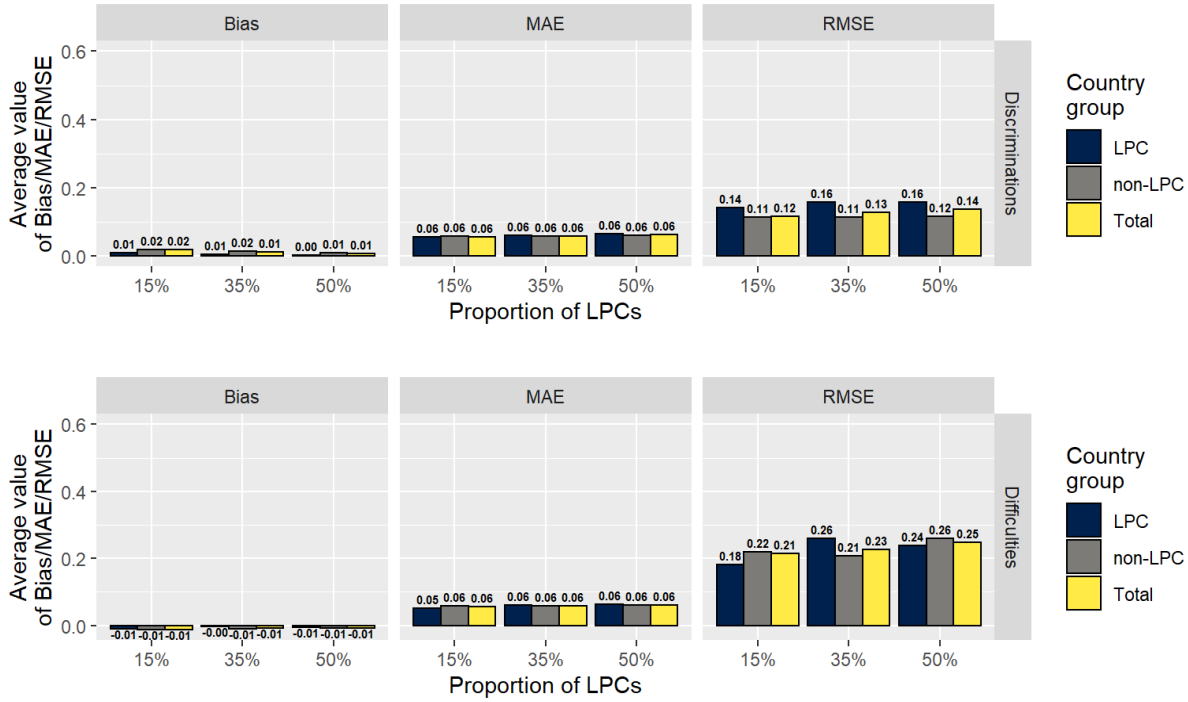
Figure 2. Bias and RMSE for country means as a function of different proportion of LPCs and countries' proficiency levels



3.1.2. Item discrimination and difficulty recovery

Item parameters are another key PISA indicators. This part of simulation tested how an increasing proportion of LPCs among PISA participants could affect recovery of item discrimination and item difficulty, two parameters estimated in models selected to scale PISA data in this study (2PLM/GPCM).

Figure 3. Recovery of item parameters for different proportion of LPCs

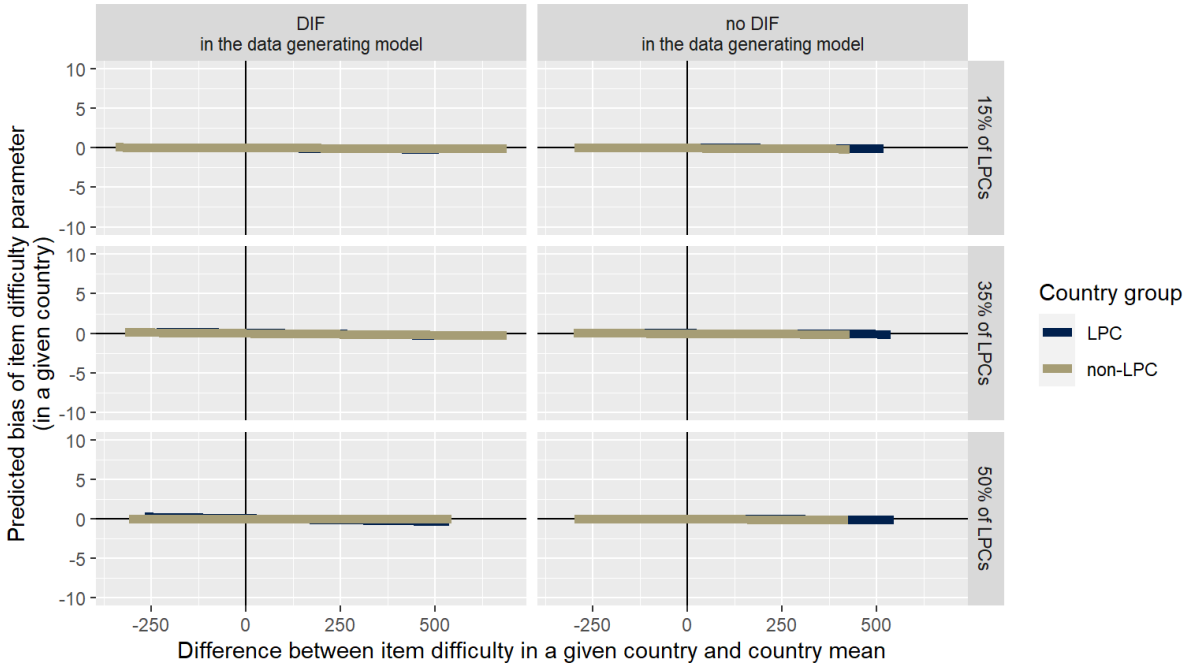


The average value of bias is very low, around 0.01, in case of both indices. The values of MAE and RMSE are also rather low. Most importantly, however, there is no indication that the proportion of LPCs among PISA participants is related to bias in item parameters recovery, which necessarily confirms the results from the analysis on country means and standard deviations as both item and group parameters are strictly related to each other.

Variation in bias of item difficulty parameters is shown in Figure 4 along with the size of a difference between item difficulty (in a given country in the data generating model) and country mean proficiency[7] (in the data generating model). There is no relationship between item difficulty and bias. This holds for both LPCs and OECD countries and irrespective of whether an item is affected by DIF (in the data generating model) or not.

[7] To make the figure clearer we provided only lines describing the size of the bias as a linear function of the size of a difference between item difficulty and country mean proficiency. These were estimated using the OLS regression.

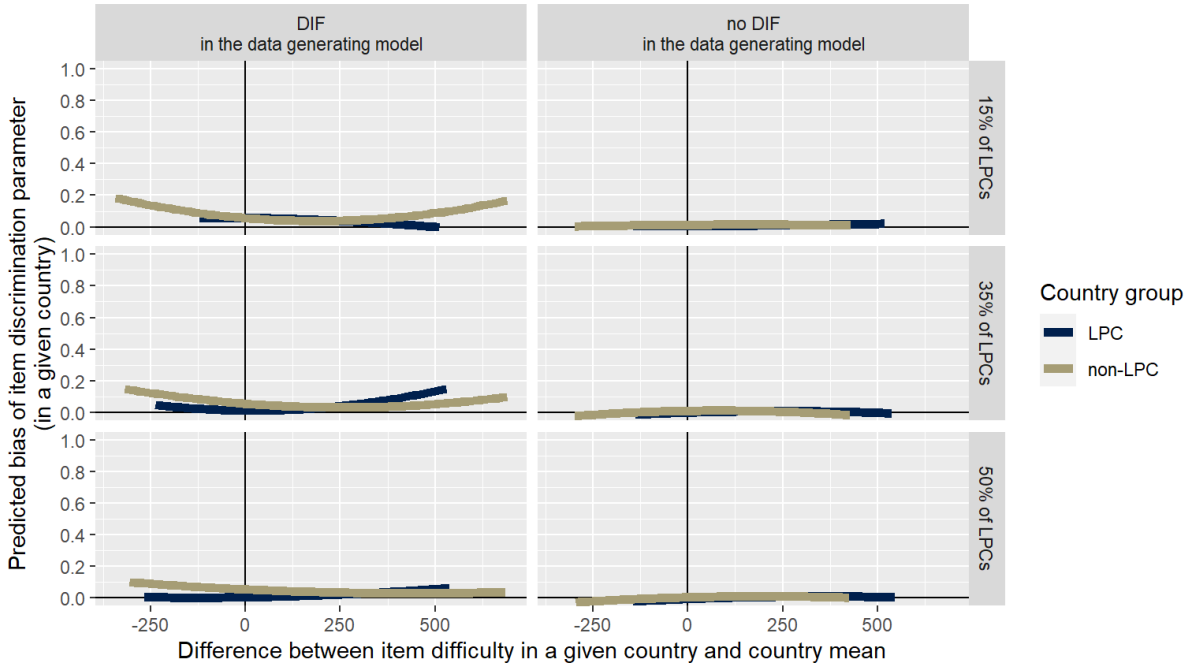
Figure 4. Bias of difficulty parameter and differences between item difficulty (in a given country in the data generating model) and country mean proficiency (in the data generating model)



Situation is somewhat more complicated with respect to bias of discrimination parameters (see Figure 5). In LPCs items affected by DIF discrimination parameters are generally the more overestimated, the larger the absolute difference between item difficulty and country mean. Similar pattern occurs also for non-LPCs but there the lowest value of bias is achieved further right from a given country mean[8]. We believe it is related to the low amount of information available in estimation of item parameters, respectively due to low variation of responses and due to the small number of students used in estimation if an item was marked as DIF and its parameters are estimated separately in a given country.

[8] To make the figure clearer we provided only lines describing the size of the bias as a quadratic function of the size of a difference between item difficulty and country mean proficiency. These were estimated using the OLS regression.

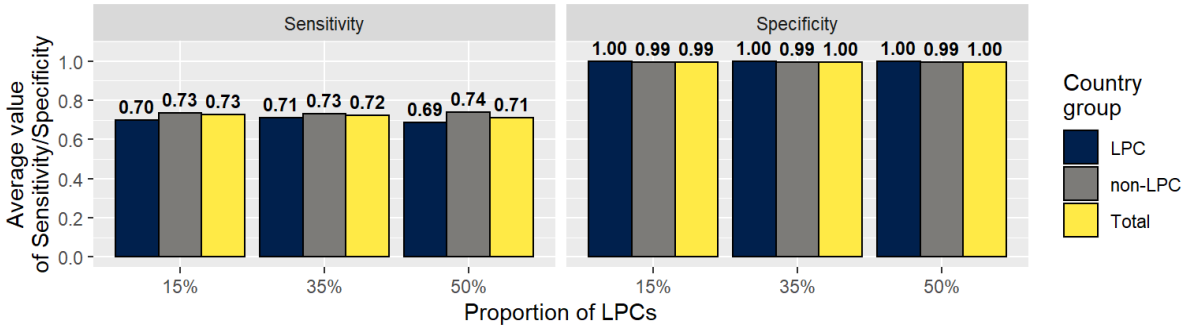
Figure 5. Bias of discrimination parameter and differences between item difficulty (in a given country in the data generating model) and country mean proficiency (in the data generating model)



3.1.3. DIF detection

Figure 6 presents the results regarding specificity and sensitivity of the DIF detection procedure. Sensitivity of DIF detection is between 73% and 74% in all presented conditions for OECD countries and between 69% and 71% for LPCs. Specificity is higher, reaching almost 100% for both groups of countries. Overall, this shows good accuracy of DIF detection, which is especially effective for non-LPCs, and only marginally worse for LPCs.

Figure 6. Sensitivity and Specificity of DIF detection for different proportion of LPCs

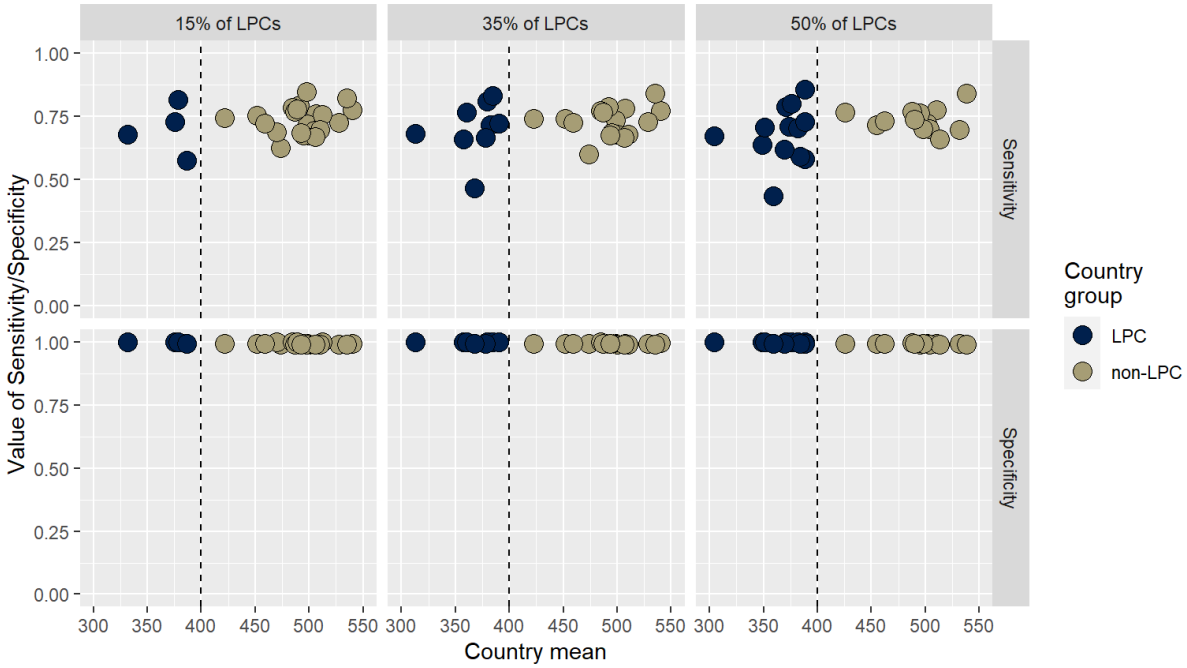


What is most important from the perspective of this report, is that neither the specificity, nor sensitivity of DIF detection does not depend on the proportion of LPCs among PISA participants. Sensitivity is rather dependent on the relative difficulty of the items to the ability distributions, something which was pointed out in previous papers e.g. (Rutkowski, Rutkowski and Liaw, 2018[4]; Tijmstra et al., 2020[8]); see also (Rutkowski and Rutkowski, 2021[6]; Rutkowski, Rutkowski and Liaw, 2019[5]). Results from the current

study are detailed in Figure 7, where sensitivity of DIF detection was plotted against country means used in this simulation study. It is evidenced that DIF detection sensitivity is slightly lower for the countries with low levels of proficiency but it is not dependent on the proportion of LPCs among PISA participants. Specificity of DIF detection, on the other hand, does not depend on the country's mean proficiency.

Slightly lower ability to detect DIF items in LPCs with very low mean proficiency confirms previous study on this topic, e.g. (Tijmstra et al., 2020[8]). However, we do not observe consequences of this fact on LPCs mean proficiency estimates that were anticipated by the cited authors. Precisely: in our study lower sensitivity does not imply underestimation of a country's mean proficiency. The most probable explanation of this phenomena is that the effect arising from the limited information provided by test items in case of students whose proficiency differs from the difficulty of items they solve (irrespective of DIF detection) results in shrinking estimates towards average difficulty of items in the item bank (in case of LPCs this means introducing positive bias to country mean estimates), cancels out the effect of omitting some undetected DIF-affected items in the model specification.

Figure 7. Sensitivity and Specificity of DIF detection for different proportion of LPCs and countries' mean proficiency



3.2. Can we do better?

In this part we investigate to what extent potential changes in the PISA assessment could affect the outcomes of the scaling. We concentrate on recovery of inter-country score differences and estimates of country means and within-country standard deviations in context of an increasing number of LPCs in PISA. Results for item parameters recovery and DIF detection could be found in the Annex.

3.2.1. Difficulty of the test

In this scenario we lower the mean difficulty of non-linking items by 100 points. This simulation condition addresses the proposition that low-performing countries should have more items that would match their ability distribution (Rutkowski, Rutkowski and Liaw, 2018[4]; 2019[5]). This scenario is relatively easy to implement and in theory should have a positive impact on the precision of PISA estimates.

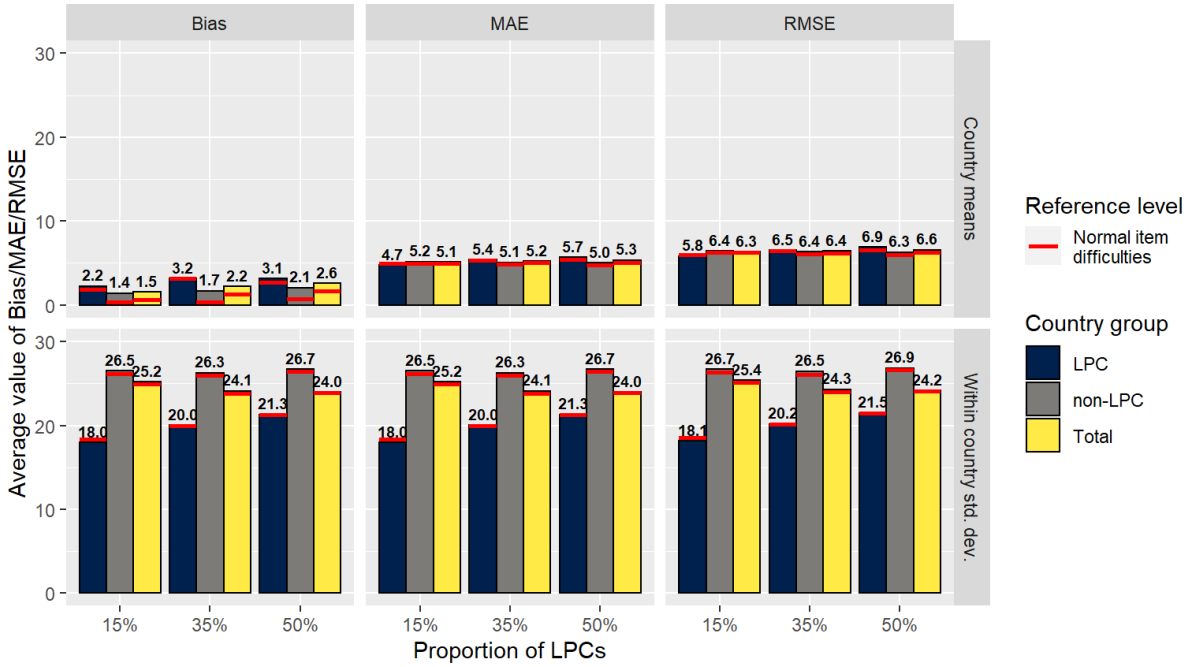
Table 2 presents correlations between true country means (that is the values used in simulating data) and estimated means by models. The values from baseline scenarios are presented in parentheses for easy interpretation of how much we could gain by introducing easier items. Lowering item difficulty hardly changes these correlations in comparison to the analysis performed on the item pool of real difficulty (as used in PISA 2015).

Table 2. Pearson's correlation between true country means and estimated country means in conditions with lower item difficulties of non-linking items. Reference values, for current type of scaling, in parentheses

	Share of LPCs		
Group characteristics	**15%**	**30%**	**50%**
Country means	0.993 (0.993)	0.996 (0.996)	0.996 (0.996)
Country standard deviations	0.915 (0.912)	0.941 (0.939)	0.948 (0.947)

Similar conclusions could be taken from Figure 8, where bias, MAE and RMSE for group means and standard deviations are presented. Here, the baseline scenario results are depicted by red horizontal bars. The bias, MAE and RMSE for country means and standard deviations are very close to what was observed in the previous analysis and this pattern holds for both groups of countries. The proportion of LPCs is not related to any of the tested indices, once again showing that key PISA indicators are robust to an increasing share of LPCs in the set of PISA participants.

Figure 8. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with lower item difficulties of non-linking items



This result also suggests that manipulating difficulty for some items in the PISA item pool is not an effective strategy to reduce bias in person and item parameters. Although in this setting we substantially lowered item difficulty, one should remember that this change refers only to the 99 non-linking items and it affects LPCs that applied computer-based assessment (roughly half of the LPCs) as paper-based assessment is composed by linking items only. Such a change in the PISA item pool in addition with a specific PISA design (rotated booklets) is simply not enough to guarantee better measurement for LPCs (Rutkowski and Rutkowski, 2021[6]; Rutkowski, Rutkowski and Liaw, 2018[4]). This shows that improving quality of measurement is a difficult task. One possible solution could be to change the difficulty of linking items but such a process needs time and may negatively impact linking accuracy across cycles. The second possible option is introducing test adaptability, what was already done in the 2018 edition of the study. This seems to be a much more promising avenue also in the future PISA cycles (Yamamoto, Shin and Khorramdel, 2018[29]), although it is not free of challenges, e.g. (Steinfeld and Robitzsch, 2021[30]).

3.2.2. Test scaling

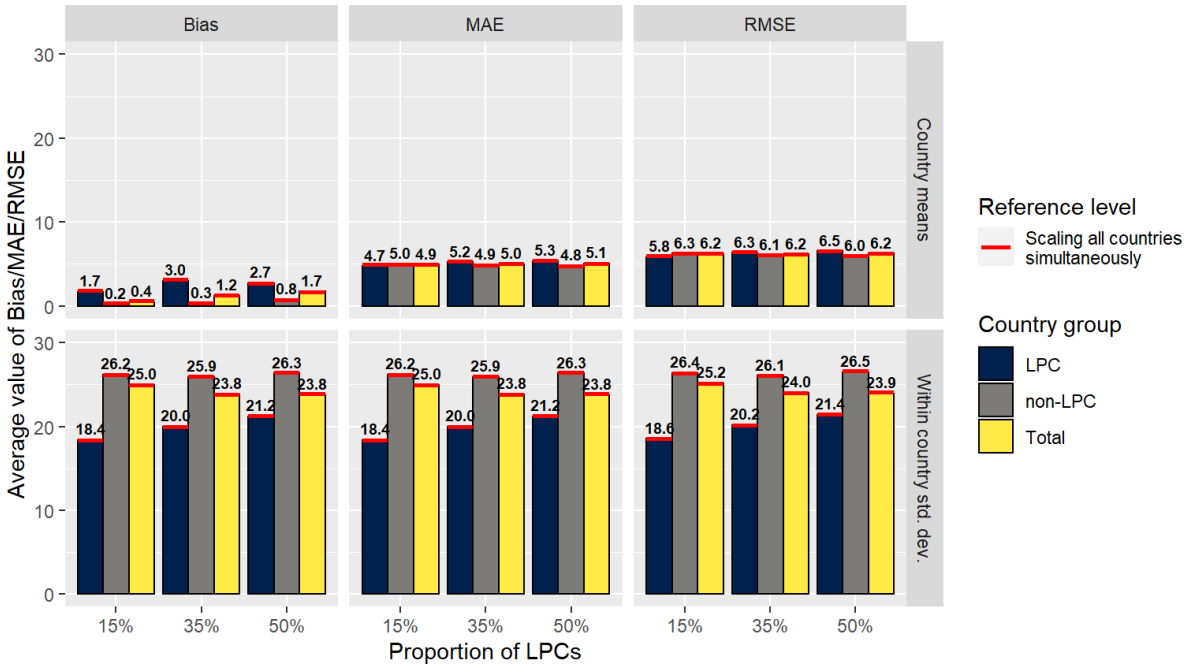
Another simulation condition tested whether scaling data from OECD countries alone would affect anyhow the values of studied parameters. The obtained results indicate that such a procedure would not lead to improvement of inter-country score differences recovery as it is presented in Table 3.

Table 3. Pearson's correlation between true country means and simulated country means in conditions with first calibration performed using only OECD countries. Reference values, for current type of scaling, in parentheses

	Share of LPCs		
Group characteristics	**15%**	**30%**	**50%**
Country means	0.993 (0.993)	0.996 (0.996)	0.996 (0.996)
Country standard deviations	0.913 (0.912)	0.940 (0.939)	0.948 (0.947)

The results presented in Figure 9 also suggest little or no reduction of bias, MAE and RMSE values. Previous research – for example (Rutkowski, Rutkowski and Zhou, 2016[19]) – showed that different ways of estimating item parameters, e.g. including different contributing countries, lead to essentially comparable results. Our results confirm this and add that there is no sign of any effect of the proportion of LPCs among PISA participants on bias, MAE and RMSE values with this kind of scaling compared to the baseline settings.

Figure 9. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with first calibration performed using only OECD countries



3.2.3. Perfect DIF detection

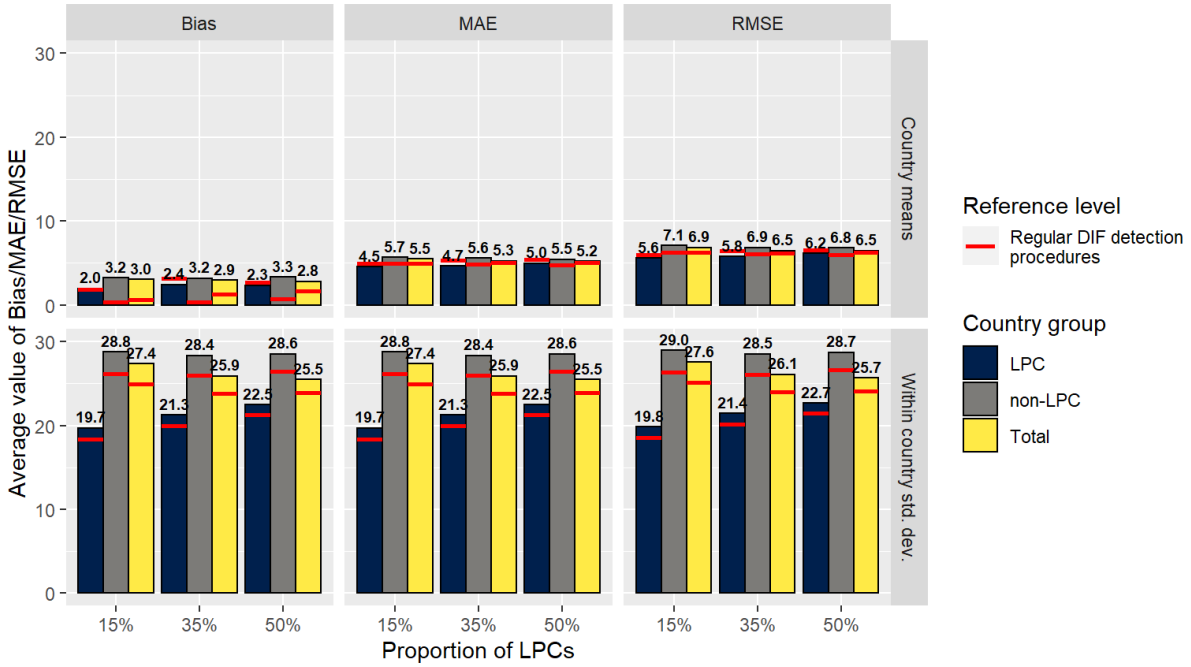
One of the most prominent and common sense assumptions is that increasing accuracy of DIF detection would result in more precise parameter estimates. However, results presented in Table 4 show that the ideal situation where all DIF items are identified (known) will not substantially improve the recovery of inter-country score differences.

Table 4. Pearson's correlation between true country means and estimated country means with perfect DIF detection. Reference values, for current type of scaling, in parentheses

	Share of LPCs		
Group characteristics	**15%**	**30%**	**50%**
Country means	0.993 (0.993)	0.996 (0.996)	0.996 (0.996)
Country standard deviations	0.916 (0.912)	0.944 (0.939)	0.950 (0.947)

More interesting results could be found in Figure 10 where bias, MAE and RMSE are presented. Regarding estimation of the country means in the OECD countries group, applying perfect DIF detection has a negative effect, that is, it slightly increases the absolute value of average bias. The difference between the baseline scenario and the scenario with perfect DIF detection is not very large, though it is noticeable and consistent across indicators and conditions with different percent of LPCs. Moreover, substantial overestimation of the within-country standard deviations was observed for the perfect DIF detection condition in both groups of countries.

Figure 10. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with Perfect DIF detection



The reasons for this result lies in technical implementation of the scaling model and sample size constraints. The more DIF is present in the model (and with perfect DIF detection the number of items that are treated as DIF items is around a quarter higher compared to the baseline scenario), the more item parameters have to be estimated freely in some countries and the overall number of model parameters grows. Moreover, DIF items that usually are not accurately detected are those with very high or very low difficulty ranges, which is followed by low variation in responses. As DIF items are estimated using only respondents from one country, the sample size is very limited (note that a given item is solved only by a small part of PISA participants from a given country due to the rotational booklet design). Combining limited samples with low variation in responses results in estimates of lower precision, especially for item slopes, which explains the poor recovery of within-country standard deviations.

Our simulation study shows therefore that the question of DIF detection is not the most important, but rather the issue of treating DIF parameters should be considered as the central one. Specifically, it is crucial not only to detect DIF but also to use good criteria whether available data enables estimating group-specific parameters of an item with a reasonable precision. If there are too few respondents and/or too low variation in responses, trying to estimate item parameters freely may cause more harm than ignoring DIF in the model specification. In operational PISA settings this problem is mitigated by grouping DIF items into clusters that are estimated together. The procedure is however hand tuned and it is difficult to replicate in a simulation study. Therefore, it was not implemented in this study.

3.2.4. Ignoring DIF

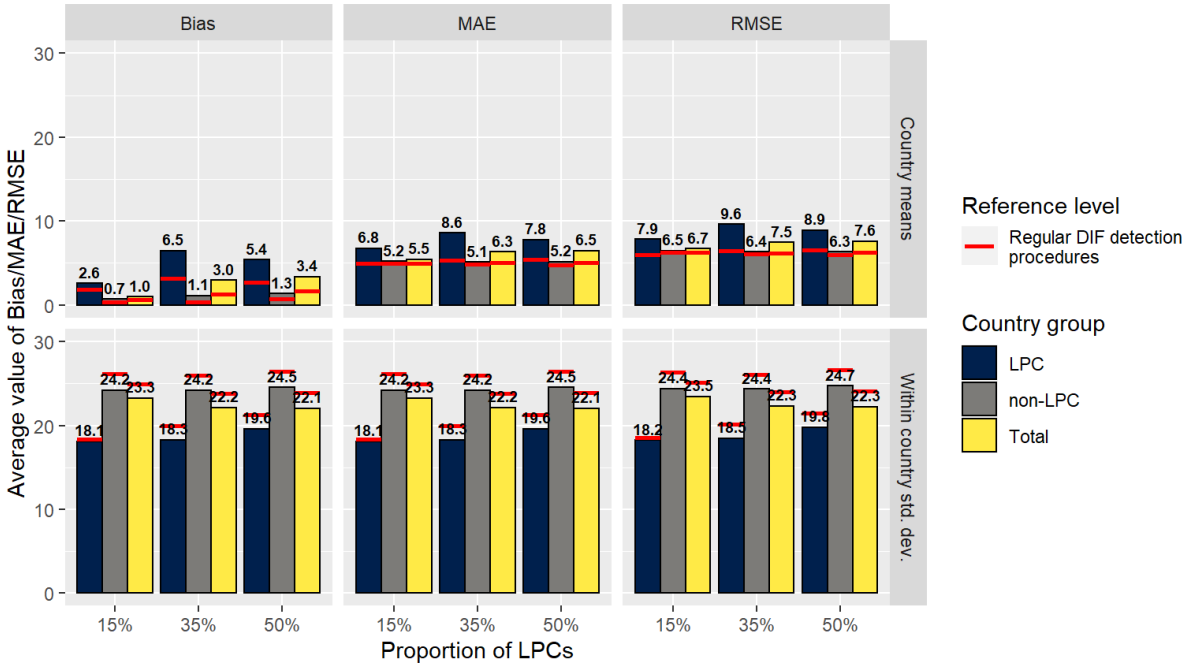
Results described in the previous section may raise a question about the overall importance of DIF detection for calibrating models based on the PISA data. Below, we answer this

question using estimates obtained from the invariant model, i.e. DIF is ignored, including both LPCs and non-LPCs (i.e. the one that was in each iteration estimated first, before applying the DIF detection procedure, note that this condition resembles the pre-2015 PISA operational procedures). Results in Table 5 indicate that ignoring DIF completely leads to a slight decrease in the recovery of inter-country score differences with respect to mean proficiency and a slight increase in recovery of differences with respect to within-country variation of proficiency.

Table 5. Pearson's correlation between true country means and estimated country means in conditions with ignoring DIF. Reference values, for current type of scaling, in parentheses

	Share of LPCs		
Group characteristics	**15%**	**30%**	**50%**
Country means	0.992 (0.993)	0.994 (0.996)	0.995 (0.996)
Country standard deviations	0.910 (0.912)	0.939 (0.939)	0.944 (0.947)

Figure 11. Recovery of country means and within-country standard deviations for different proportion of LPCs among PISA participants in conditions with ignoring DIF (calibrating only the invariant model)



In conclusion, while applying DIF detection procedures does not seem so important when considering simple country rankings, it gains importance when turning to more detailed analysis of the PISA data.

4. Discussion

The main motivation for this research was concern about an increasing proportion of LPCs in PISA and its effect on the measurement quality of PISA-based indicators. The results obtained in our study enable us to answer this question clearly: the scaling methodology applied in PISA provides robust estimates of inter-country score differences regardless of the proportion of LPCs among PISA participants. Moreover, expanding PISA on more LPCs should have rather positive effects as adding more countries increases the sample size for item estimation and allows for better estimates of item parameters, especially in case of easy tasks. If easy items are not burdened with DIF then this will allow for more precise estimates of the skills of the lower performing students both from LPCs and from OECD countries and, consequently, better estimates for the whole set of PISA participants.

Moreover, there are no straightforward methods that could improve the methodology of scaling PISA data: neither separate scaling of OECD countries, nor introducing easier items is a feasible method of improving PISA estimates. Our study points out further developments of PISA methodology that could alleviate the detected problems.

In our opinion not only DIF detection procedures (investigated by many authors) but also DIF handling procedures (much less researched) should be investigated in more detail. Specifically, results of our analysis show that with low variation of responses to the item in a given group (what happens if item difficulty differs largely from a given group mean proficiency) abandoning DIF modelling may be more beneficial than trying to estimate

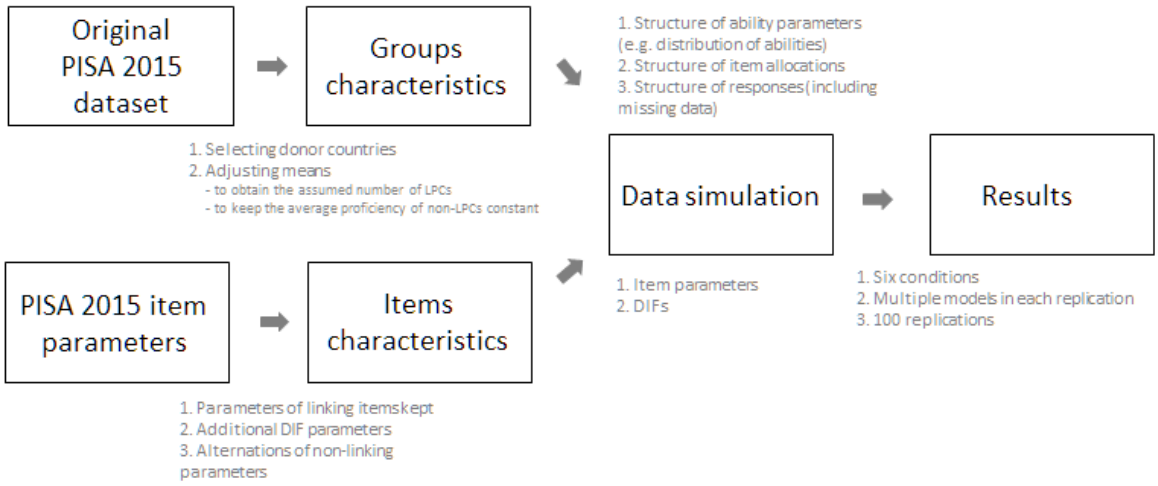
item parameters separately for this group. Consequently, lower ability to detect DIF in LPCs using fixed threshold of the RMSD statistic, documented in a recent publication (Tijmstra et al., 2020[8]) should be seen as having limited practical importance. This is because many DIFs, even if detected, in many cases cannot be reasonably modelled anyway, due to insufficient information in the data to assure precise estimation of item parameters in a given group. Presented problems are specific for the concurrent scaling based on partial invariance. Some works show that we could also consider other ways of scaling e.g. robust linking approaches (Robitzsch and Lüdtke, 2020[31]) or alignment optimisation (Muthén and Asparouhov, 2014[24]; Pokropek, Lüdtke and Robitzsch, 2020[32]). Robitzsch and Lüdtke (2020[31]) showed that the robust linking approaches (so-called Robust Haberman and Robust Haebara) performed similarly to the partial invariance approach using the RMSD in rather artificial conditions (small sample size [250, 500, 1 000], low number of countries [20] and narrow range of country means and standard deviations). Also Pokropek, Lüdtke and Robitzsch (2020[32]) showed promising results for generalised alignment approach but once again with a limited number of groups (8) and small number of items (20). However, alternative methods seem promising and it would be interesting to investigate its efficacy in conditions more similar to real life settings where partial invariance faces significant problems.

5. Method in details

5.1. Data generation procedures

The overall structure of the simulation design is presented in Figure 12. We start with original PISA 2015 data to exactly mimic the structure of the assessment. Therefore, group characteristics used in the simulation imitate the real data situation as far as it is possible. Item parameters used for the simulation were also taken from real data; some of them were kept as estimated in PISA 2015 study, some were changed, depending on simulation conditions.

Figure 12. Structure of the simulation design



For each condition we performed 100 replications using estimation procedures closely resembling PISA scaling. Below, we describe details of the data generation, simulation, and estimation procedures.

5.1.1. Group characteristics

In order to bring the simulation conditions as close as possible to PISA operational procedures, group (country) characteristics are based on real PISA 2015 data. Similar method was used by Rutkowski and colleagues (2018; 2019), where country means and standard deviations were borrowed from real countries to simulate abilities distribution. In our approach we go one step further and model the full test design, sample design and respondents' missing data patterns as they are in the real PISA implementation. The countries from which we obtain information are referred to as donor countries. The simulated groups reflecting properties of countries are called recipients. Donor countries were selected from PISA 2015 data. To reduce complexity, we selected countries where the majority of students use one testing language (at least 70% of participants sitting for an assessment in the dominant language). Only data and characteristics of this dominant language group were used for each country.

Table 6. Countries used as donors for the simulation

Donor countries									Means used in the data generating model in a given condition		
CNT	Lang.	OECD 2015	Mode[9]	Mean	SD	ICC	N	Cluster	1 (22+4)	2 (17+9)	3 (2x13)
DOM	Spanish	No	c	331,6	72,5	0,41	4 740	LPC	331,6	313,1	304,5
DZA	Arabic	No	p	375,7	69,3	0,34	5 519	LPC	375,7	357,3	348,6
KSV	Albanian	No	p	378,4	71,3	0,33	4 826	LPC	378,4	360,0	351,3
TUN	Arabic	No	c	386,4	64,9	0,39	5 375	LPC	386,4	367,9	359,3
PER	Spanish	No	c	396,7	76,7	0,40	6 971	LPC	ns	378,2	369,6
MKD	Macedonian	No	p	397,6	83,8	0,27	3 895	LPC	ns	379,1	370,5
BRA	Portuguese	No	c	400,7	89,2	0,42	23 141	LPC	ns	382,2	373,6
IDN	Bahasa	No	p	403,1	68,4	0,44	6 513	LPC	ns	384,6	376,0
JOR	Arabic	No	p	408,7	84,4	0,30	7 267	LPC	ns	390,2	381,6
MNE	Serb	No	c	411,3	85,3	0,27	5 665	LPC	ns	ns	384,2
GEO	Georgian	No	p	415,4	89,0	0,23	4 954	LPC	ns	ns	388,3
MEX	Spanish	Yes	c	415,7	71,4	0,34	7 568	LPC	ns	ns	388,6
COL	Spanish	No	c	415,7	80,4	0,36	11 795	LPC	ns	ns	388,6
TUR	Turkish	Yes	c	425,5	79,3	0,55	5 895	Non-LPC	421,9	422,4	425,5
GRC	Greek	Yes	c	454,8	91,9	0,40	5 532	Non-LPC	451,3	451,7	454,8
SVK	Slovak	Yes	c	462,3	98,8	0,46	5 948	Non-LPC	458,8	459,2	462,3
ISL	Icelandic	Yes	c	473,2	91,2	0,08	3 371	Non-LPC	469,7	ns	ns
HUN	Hungarian	Yes	c	476,7	96,3	0,58	5 658	Non-LPC	473,2	473,6	ns
LVA	Latvian	Yes	c	487,6	82,5	0,24	3 584	Non-LPC	484,1	484,5	487,6
ESP	Spanish	Yes	c	490,0	87,2	0,16	5 092	Non-LPC	486,4	486,8	490,0
SWE	Swedish	Yes	c	492,5	102,2	0,19	5 387	Non-LPC	488,9	ns	ns
FRA	French	Yes	c	495,0	102,0	0,52	6 108	Non-LPC	491,4	491,9	495,0
USA	English	Yes	c	496,2	98,6	0,22	5 712	Non-LPC	492,7	493,1	ns
NOR	Bokmal	Yes	c	498,1	96,3	0,12	5 007	Non-LPC	494,5	495,0	498,1
POL	Polish	Yes	c	501,4	90,8	0,18	4 478	Non-LPC	497,9	ns	ns
DNK	Danish	Yes	c	501,9	90,3	0,19	7 161	Non-LPC	498,4	498,8	501,9
IRL	English	Yes	c	503,4	88,7	0,16	5 638	Non-LPC	499,9	500,3	503,5
NLD	Dutch	Yes	c	508,6	101,0	0,59	5 385	Non-LPC	505,0	ns	ns
GBR	English	Yes	c	509,4	99,7	0,25	13 818	Non-LPC	505,8	506,3	ns
AUS	English	Yes	c	510,0	102,3	0,26	14 530	Non-LPC	506,4	506,9	510,0
NZL	English	Yes	c	513,3	104,1	0,22	4 520	Non-LPC	509,7	510,2	513,3
KOR	Korean	Yes	c	515,8	95,2	0,27	5 581	Non-LPC	512,2	ns	ns
FIN	Finnish	Yes	c	531,2	96,3	0,13	5 534	Non-LPC	527,6	528,1	531,2
JPN	Japanese	Yes	c	538,4	93,5	0,46	6 647	Non-LPC	534,8	535,3	538,4
EST	Estonian	Yes	c	543,5	87,7	0,19	4 338	Non-LPC	539,9	540,4	ns

[9] Note: Mode: c - computer; p - paper; ns - not selected

For data generation procedures countries were divided into two clusters – non-LPC and LPC countries (countries with an average PISA science score below 400 points). In PISA 2015, our reference dataset, the number of LPCs with one dominant language in 2015 was limited to only six countries. This is sufficient only for the first condition defined by the proportion of LPCs (22+4), but not for the remaining two (17+9 and 13+13). Therefore, we decided to use donors with means of up to 416 PISA points with their means linearly shifted downwards, so that all means in the LPC cluster were below 400 PISA points and that the average of LPCs' means was 368 PISA points, which corresponds to the mean level of PISA points in the LPC group in the first condition. The donor countries and their group level parameters used for simulations are displayed in Table 6 below. The last three columns present the group means used for simulations. This procedure yields all LPCs having a mean below 400 PISA points. Means of OECD countries were also shifted to equalise the OECD average in all conditions.

Please note, that this procedure results in means of simulated countries different from the original means of donor countries and that the simulated means differ between the conditions by design. Country means were adjusted in each of the simulation conditions in order to achieve a twofold goal: a) to obtain the assumed number of LPCs (countries with mean proficiency below 400 PISA points), b) to keep the average proficiency of non LPCs constant across conditions.

5.1.2. Item characteristics

For the simulation we used one domain, basing on the fact that starting from the newest PISA cycles (2015 and 2018) the number of items in major and minor domains became similar. We used 184 items, as in the PISA 2015 major domain's main survey (OECD, 2017[20]).

IRT parameters of PISA 2015 science items were specified using publicly available information about PISA 2015 trend item parameters; see Annex H of (OECD, 2017[20]). Parameters for the rest of the items (including new items, designed specifically for PISA 2015) were estimated using a partially-invariant one-dimensional model in the form that was used to scale PISA 2015 data with item parameters enabled to vary for some groups, according to the DIF patterns described in PISA 2015 documentation – see Annex G of (OECD, 2017[20]) – and parameters of trend items fixed to values from the original PISA 2015 study; see Annex H of (OECD, 2017[20]).

In the case of two items, PS519Q02S and PS413Q04S for Republic of Northern Macedonia, which parameters were consequently estimated freely for this country, we assessed the estimated difficulty parameters as unreliable: -4.9 and 7.72, respectively. The values of corresponding parameters in a non-varying country cluster were -0.20 and 0.25, respectively. Moreover, these items' discrimination parameters in FYROM were very low. Therefore, we decided to override these estimated values with the difficulties estimated for a non-varying country cluster to deal with the problem of unrealistic parameters.

Moreover, it was randomly determined, which items in each of the countries are affected by *additional* DIF, namely DIF that is assumed to remain undetected by the currently employed DIF detection method (see simulation conditions for details). DIF is generated as in the real PISA data, hence we assume presence of both uniform and non-uniform DIF. Each generated group reflects the DIF structure regarding size and number of items affected in a particular country selected from PISA 2015 (OECD, 2017[20]). The number of affected items (that is equal to the unique item parameters reported in PISA) is increased by approximately 30% for the OECD countries and 60% for the LPCs. The primary aim of this simulation is to investigate the effect of increasing participation of LPCs in PISA study

and not (at least not directly) the increase of DIF items nor DIF size. The increase of DIF items is an attempt to make the design of the simulations as close to the real life scenario as possible. Previous simulation studies (Finch, 2016[33]; Lin, 2020[34]; Pokropek and Pokropek, 2021[35]) suggest that various DIF detection methods in conditions close to those encountered for PISA OECD countries reach between 50% and 75% of true positive rate (*sensitivity*, correct DIF detection). Moreover, additional studies (Rutkowski and Rutkowski, 2021[6]; Tijmstra et al., 2020[8]) provide evidence that the true positive rate of DIF for LPCs is substantially lower, therefore more DIF-affected items will be added for these countries, as we believe that a moderate increase of the number of DIF items in the generating model compared to the detected DIF items in the standard PISA analysis is a rational decision that makes our design more plausible. The size of additional DIF effects mimic small (0.4), medium (0.6), and large (0.8) effect sizes for uniform DIF (Finch, 2016[33]; Penfield, 2001[36]) and 0.6 for non-uniform DIF (Holmes Finch and French, 2007[37]; Woods, 2008[38]). The direction of additional DIF was selected at random independently of the DIF size and countries characteristics. The direction of uniform and the direction of non-uniform DIF for the same item also was sampled independently of each other.

5.1.3. Data simulation

For each of the selected donor countries 5 000 participants were sampled independently in each iteration of the simulation to provide patterns of school allocation (respondents nested in schools), booklet assignment, and missing data.

To make simulation as similar to the real PISA design as possible, we assume a sample size of 5 000 (i.e. the sum of *senate* weights within each country) and we also imitate PISA complex sampling design and the incomplete block design, namely how students are drawn to PISA and how items are assigned to students.

Use of senate weights is mimicked by using an equal number of observations in each country sampled with the probability proportional to the final weights in the original sample. In this way we achieve two goals: a) each country contributes equally while estimating item parameters and b) simulation design accounts for unequal probability sampling.

Testlets/units were designed as in PISA 2015 main domain (science). Here we represent how items were embedded into questions (units), which consisted of a stem (text, image, introduction) and a number of items attached to it. We assume no violations of the assumption of the conditional independence of item responses (i.e. that students respond in line with the IRT model that is used to scale the PISA data, up to the possibility of some undetected or falsely detected DIF in some country-language groups), all using the procedures exactly as in the real PISA 2015 study.

In each iteration parameters describing PISA 2015 major domain (science) achievements in these countries - means, standard deviations and intra-class correlations (ICCs) on the school level - were used to generate proficiency estimates for 5 000 observations in each country in a two-stage procedure imitating complex survey design used in PISA. First, school-level means were generated for the number of schools equal to the number of distinct schools of the sampled participants. Individual achievements were generated afterwards as deviations from these means.

We assume that the true population model is the Two Parameter Logistic Model or the Generalized Partial Credit IRT Model (2PLM/GPCM). These models are operationally used in scaling PISA 2015 data and proven to be relatively well-fitted to the data (OECD, 2017[20]). Using item parameters, knowledge on DIF patterns, participants' achievements

(abilities), and missing data patterns (either due to design or due to item omission), responses to items were generated in each iteration according to the one-dimensional 2PL/GPC model.

Within each country the previously sampled patterns of booklet assignment and missing data were assigned to the generated observations by means of achievements (i.e. generated observations were ordered within each country with respect to their achievements and patterns previously sampled from the PISA 2015 dataset were ordered according to estimated achievements of participants they represent and then matched together). This procedure enables to fully mimic booklet assignment and missing data patterns from the PISA 2015 main study. This procedure enables mimicking the booklet-to-students assignment also in adaptive test design, including mimicking the strength of adaptiveness - the degree of relation between booklet assignment and trait level is exactly the same as in the original study. Hence, if this relation is zero in the original dataset, it is also zero in the generated data.

5.2. Estimation procedures

Estimation procedures were the same as original PISA operational scaling procedures (OECD, 2017[20]). After generating item responses four IRT models were estimated:

1. partially-invariant, one-dimensional 2PL/GPC model with DIF-affected item group pairs specified according to information about the data generating process, i.e. as if perfect DIF detection procedure was available
2. fully-invariant, one-dimensional 2PL/GPC model in the form that is used in the first step of PISA item calibration procedure
3. two models aimed to diminish the potential impact of DIF in LPCs on recovery of item parameters in non-LPCs:
    a. partially-invariant, one-dimensional 2PL/GPC model specified according to information about DIF in the linking items set (assuming all the other items being invariant) was estimated only on data from the non-LPC cluster
    b. partially-invariant, one-dimensional 2PL/GPC model was estimated on data from both country clusters (LPC and OECD), fixing all item parameters on values estimated using only non-LPC data.

Next, separately for models described in points 2 and 3.b above, we followed the current PISA DIF detection procedure (OECD, 2017[20]): for each item in each group the RMSD (root mean square deviation) statistic was computed using results from the model described in point 2 or 3b. Items for which the RMSD statistic exceeded the assumed threshold, were marked as DIF, yielding group non-invariance, namely differences in item parameters between OECD and LPCs. Next, model specification was updated to include the detected DIF and estimation was performed again. Procedure was continued until the RMSD statistics fell below the threshold for all the items in all the groups. Mimicking real PISA scaling procedures, we started from a higher value of the threshold (0.30 and 0.20) to avoid false positives in DIF detection and only then gradually lowered it to the threshold used in PISA (0.12); see p. 151 of (OECD, 2017[20]). For each of these values the procedure was continued until no additional DIF was detected. Then, the value of the RMSD threshold was lowered and the procedure was continued. To speed up computations in cases where RMSD statistics just above the threshold of 0.12 were still found after many consecutive estimations, we stopped the procedure if the value of the threshold was already set to 0.12 and 10 steps (model estimations) of DIF detection were already performed. The only important simplification of the procedure compared to this used in PISA was that item

parameters were estimated separately in each group in which RMSD statistic exceeded the threshold. In operational procedure all the groups in which DIF for a given item appears to have the same direction are supposed to have a common value of the parameter in the next calibration (however, inspection of the RMSD statistic after this next calibration may lead to further splits). The procedure is, however, hand tuned and difficult to replicate in a simulation study and therefore it was not implemented in this study.

Parameters of the last model estimated in the previous step were used to investigate recovery (bias and accuracy) of the data generating model parameters. For estimation we used the R packages TAM version 3.7-16 (Test Analysis Models) (Robitzsch, Kiefer and Wu, 2022[39]) and Rstyles version 0.3.0 (Żółtak, Pokropek and Muszyński, 2021[40]). We used an integration grid of 41 points, equally spaced between -4 and 4, i.e. the same as used by default in PISA 2015 operational procedures for IRT model estimation using TAM. For each condition we performed 200 replications. Although the number of replications may not seem very large, we were forced to restrict our simulation due to a very demanding estimation procedure used in this study. This number of replications is higher than in many previous complex simulation studies, (Kim et al., 2017[41]; Meade and Lautenschlager, 2004[42]; Nylund, Asparouhov and Muthén, 2007[43]; Rutkowski, Rutkowski and Zhou, 2016[19]).

References

Finch, W. (2016), "Detection of Differential Item Functioning for More Than Two Groups: A Monte Carlo Comparison of Methods", *Applied Measurement in Education*, Vol. 29/1, pp. 30-45, https://doi.org/10.1080/08957347.2015.1102916. [33]

Guenole, N. and A. Brown (2014), "The consequences of ignoring measurement invariance for path coefficients in structural equation models", *Frontiers in Psychology*, Vol. 5, https://doi.org/10.3389/fpsyg.2014.00980. [9]

Holmes Finch, W. and B. French (2007), "Detection of Crossing Differential Item Functioning", *Educational and Psychological Measurement*, Vol. 67/4, pp. 565-582, https://doi.org/10.1177/0013164406296975. [37]

Kamens, D. and C. McNeely (2010), "Globalization and the Growth of International Educational Testing and National Assessment", *Comparative Education Review*, Vol. 54/1, pp. 5-25, https://doi.org/10.1086/648471. [1]

Khorramdel, L., H. Shin and M. von Davier (2019), "GDM Software mdltm Including Parallel EM Algorithm", in *Handbook of Diagnostic Classification Models*, https://doi.org/10.1007/978-3-030-05584-4_30. [7]

Kim, E. et al. (2017), "Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 24/4, pp. 524-544, https://doi.org/10.1080/10705511.2017.1304822. [41]

Köhler, C. and J. Hartig (2017), "Practical Significance of Item Misfit in Educational Assessments", *Applied Psychological Measurement*, Vol. 41/5, pp. 388-400, https://doi.org/10.1177/0146621617692978. [10]

L. Khorramdel, M. (ed.) (Forthcoming), *Linking International Large-Scale Assessments Over Time and Administration Modes*, Springer. [21]

Lin, L. (2020), *Evaluate Measurement Invariance Across Multiple Groups: A Comparison Between The Alignment Optimization And The Random Item Effects Model*, University of Pittsburgh, Pittsburgh, http://d-scholarship.pitt.edu/id/eprint/39977 (accessed on 13 July 2022). [34]

Lockheed, M., T. Prokic-Bruer and A. Shadrova (2015), *The Experience of Middle-Income Countries Participating in PISA 2000-2015*, PISA, The World Bank, Washington, D.C./OECD Publishing, Paris, https://doi.org/10.1787/9789264246195-en. [2]

Marsh, H. et al. (2018), "What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups.", *Psychological Methods*, Vol. 23/3, pp. 524-545, https://doi.org/10.1037/met0000113. [27]

Meade, A. and G. Lautenschlager (2004), "A Monte-Carlo Study of Confirmatory Factor Analytic Tests of Measurement Equivalence/Invariance", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 11/1, pp. 60-72, https://doi.org/10.1207/S15328007SEM1101_5. [42]

Mislevy, R. (1991), "Randomization-based inference about latent variables from complex samples", *Psychometrika*, Vol. 56/2, https://doi.org/10.1007/BF02294457. [13]

Mislevy, R. and K. Sheeham (1987), "Marginal estimation procedures", in Beaton, A. (ed.), *Implementing the New Design: The NAEP 1983-84 Technical Teport (Report No. 15-TR-20)*, National Assessment of Educational Progress, Princeton NJ, https://files.eric.ed.gov/fulltext/ED288887.pdf (accessed on 13 July 2022). [14]

Muthén, B. and T. Asparouhov (2018), "Recent Methods for the Study of Measurement Invariance With Many Groups", *Sociological Methods & Research*, Vol. 47/4, pp. 637-664, https://doi.org/10.1177/0049124117701488. [25]

Muthén, B. and T. Asparouhov (2014), "IRT studies of many groups: the alignment method", *Frontiers in Psychology*, Vol. 5, https://doi.org/10.3389/fpsyg.2014.00978. [24]

Nylund, K., T. Asparouhov and B. Muthén (2007), "Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 14/4, pp. 535-569, https://doi.org/10.1080/10705510701575396. [43]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf (accessed on 13 July 2022). [20]

OECD (2014), *PISA 2012 Technical Report*, OECD Publishing, Paris, https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf (accessed on 13 July 2022). [23]

Oliveri, M. and M. Von Davier (2011), "Investigation of model fit and score scale comparability in international assessments", *Psychological Test and Assessment Modeling*, Vol. 53, pp. 315-333, https://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf (accessed on 13 July 2022). [11]

Penfield, R. (2001), "Assessing Differential Item Functioning Among Multiple Groups: A Comparison of Three Mantel-Haenszel Procedures", *Applied Measurement in Education*, Vol. 14/3, pp. 235-259, https://doi.org/10.1207/S15324818AME1403_3. [36]

Pokropek, A., E. Davidov and P. Schmidt (2019), "A Monte Carlo Simulation Study to Assess The Appropriateness of Traditional and Newer Approaches to Test for Measurement Invariance", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 26/5, pp. 724-744, https://doi.org/10.1080/10705511.2018.1561293. [26]

Pokropek, A., O. Lüdtke and A. Robitzsch (2020), "An extension of the invariance alignment method for scale linking", *Psychological Test and Assessment Modeling*, Vol. 62, pp. 305-334, https://www.researchgate.net/publication/342182874_An_extension_of_the_invariance_alignment_method_for_scale_linking (accessed on 13 July 2022). [32]

Pokropek, A. and E. Pokropek (2021), "Deep Neural Networks for Detecting Statistical Model Misspecifications. The Case of Measurement Invariance", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 29/3, pp. 394-41, https://doi.org/10.1080/10705511.2021.2010083. [35]

Robitzsch, A., T. Kiefer and M. Wu (2022), *TAM: Test Analysis Modules*, https://CRAN.R-project.org/package=TAM (accessed on 13 July 2022). [39]

Robitzsch, A. and O. Lüdtke (2020), "A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments", *Psychological Test and Assessment Modeling*, Vol. 62, pp. 233-279, https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-2/03_Robitzsch.pdf (accessed on 13 July 2022). [31]

Robitzsch, A. et al. (2020), "Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects", *Frontiers in Psychology*, Vol. 11, https://doi.org/10.3389/fpsyg.2020.00884. [22]

Rolfe, V. (2021), "Tailoring a measurement model of socioeconomic status: Applying the alignment optimization method to 15 years of PISA", *International Journal of Educational Research*, Vol. 106, p. 101723, https://doi.org/10.1016/j.ijer.2020.101723. [28]

Rutkowski, D. and L. Rutkowski (2021), "Running the Wrong Race? The Case of PISA for Development", *Comparative Education Review*, Vol. 65/1, pp. 147-165, https://doi.org/10.1086/712409. [6]

Rutkowski, D., L. Rutkowski and Y. Liaw (2018), "Measuring Widening Proficiency Differences in International Assessments: Are Current Approaches Enough?", *Educational Measurement: Issues and Practice*, Vol. 37/4, pp. 40-48, https://doi.org/10.1111/emip.12225. [4]

Rutkowski, L. and D. Rutkowski (2018), "Improving the Comparability and Local Usefulness of International Assessments: A Look Back and A Way Forward", *Scandinavian Journal of Educational Research*, Vol. 62/3, pp. 354-367, https://doi.org/10.1080/00313831.2016.1261044. [3]

Rutkowski, L., D. Rutkowski and Y. Liaw (2019), "The existence and impact of floor effects for low-performing PISA participants", *Assessment in Education: Principles, Policy & Practice*, Vol. 26/6, pp. 643-664, https://doi.org/10.1080/0969594X.2019.1577219. [5]

Rutkowski, L., D. Rutkowski and Y. Zhou (2016), "Item Calibration Samples and the Stability of Achievement Estimates and System Rankings: Another Look at the PISA Model", *International Journal of Testing*, Vol. 16/1, pp. 1-20, https://doi.org/10.1080/15305058.2015.1036163. [19]

Sinharay, S. and S. Haberman (2014), "How Often Is the Misfit of Item Response Theory Models Practically Significant?", *Educational Measurement: Issues and Practice*, Vol. 33/1, pp. 23-35, https://doi.org/10.1111/emip.12024. [12]

Steinfeld, J. and A. Robitzsch (2021), "Conditional Maximum Likelihood Estimation in Probability-Branched Multistage Designs", https://doi.org/10.31234/osf.io/ew27f. [30]

Thomas, N. (2002), "The role of secondary covariates when estimating latent trait population distributions", *Psychometrika*, Vol. 67/1, pp. 33-48, https://doi.org/10.1007/BF02294708. [15]

Tijmstra, J. et al. (2020), "Sensitivity of the RMSD for Detecting Item-Level Misfit in Low-Performing Countries", *Journal of Educational Measurement*, Vol. 57/4, https://doi.org/10.1111/jedm.12263. [8]

Von Davier, M., E. Gonzalez and R. Mislevy (2009), "What are plausible values and why are they useful?", in *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments: Volume 2*, IEA-ETS Research Institute, https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02.pdf (accessed on 13 July 2022). [17]

von Davier, M. et al. (2006), "32 The Statistical Procedures Used in National Assessment of Educational Progress: Recent Developments and Future Directions", *Handbook of Statistics*, Vol. 26, pp. 1039-1055, https://doi.org/10.1016/S0169-7161(06)26032-2. [16]

Woods, C. (2008), "Likelihood-Ratio DIF Testing: Effects of Nonnormality", *Applied Psychological Measurement*, Vol. 32/7, pp. 511-526, https://doi.org/10.1177/0146621607310402. [38]

Yamamoto, K., H. Shin and L. Khorramdel (2018), "Multistage Adaptive Testing Design in International Large-Scale Assessments", *Educational Measurement: Issues and Practice*, Vol. 37/4, pp. 16-27, https://doi.org/10.1111/emip.12226. [29]

Zieger, L. et al. (2020), "Conditioning: How background variables can influence PISA scores", *CEPEO Working Paper Series*, No. No 20-09, UCL Centre for Education Policy and Equalising Opportunities, https://EconPapers.repec.org/RePEc:ucl:cepeow:20-09 (accessed on 13 July 2022). [18]

Żółtak, T., A. Pokropek and M. Muszyński (2021), *tzoltak/rstyles: Version 0.4.0*, Zenodo, https://doi.org/10.5281/zenodo.5175326. [40]

Annex A. Additional graphs for simulation results and simulation code

Additional graphs for 3.1

Figure A A.1. MAE for country means for different proportion of LPCs and countries' mean proficiency

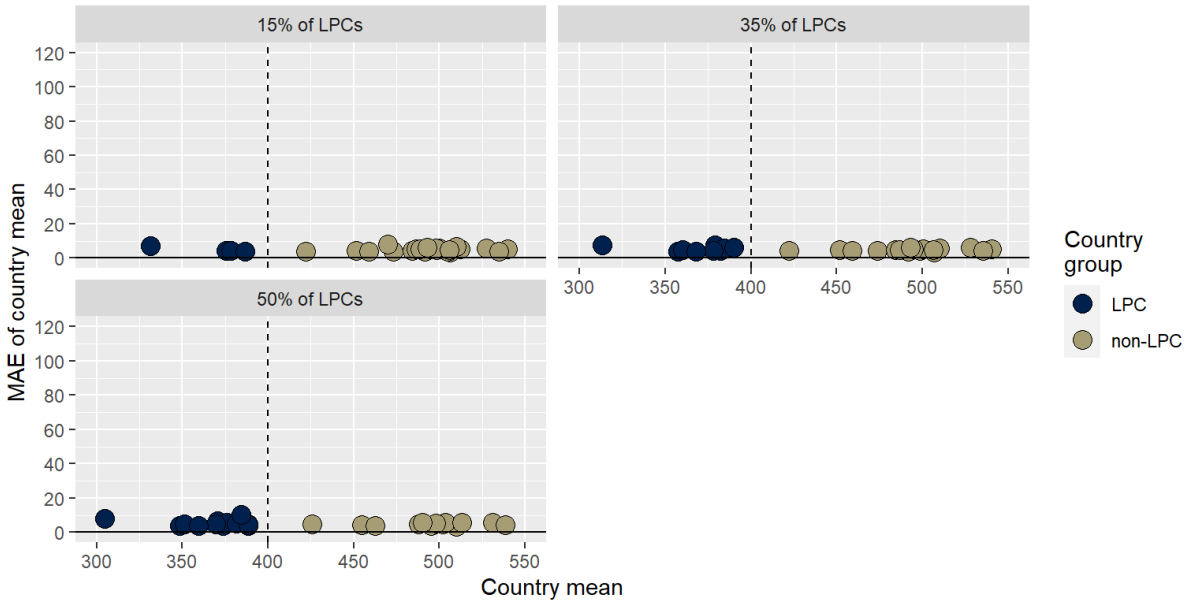


Figure A A.2. Bias for country standard deviations for different proportion of LPCs and countries' mean proficiency

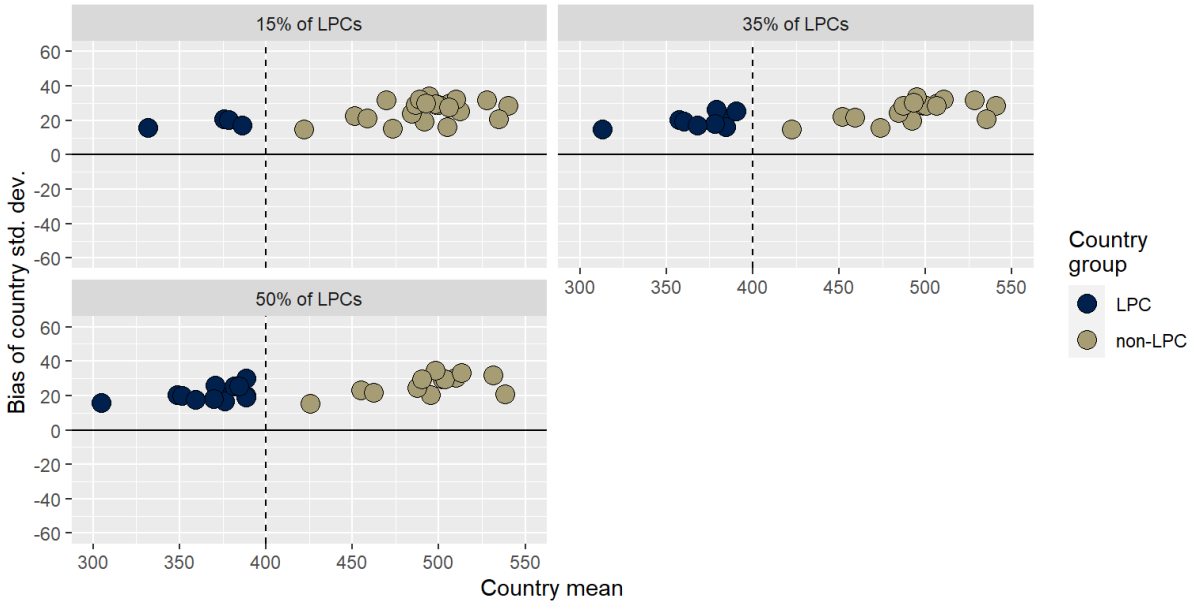


Figure A A.3. MAE for country standard deviations for different proportion of LPCs and countries' mean proficiency

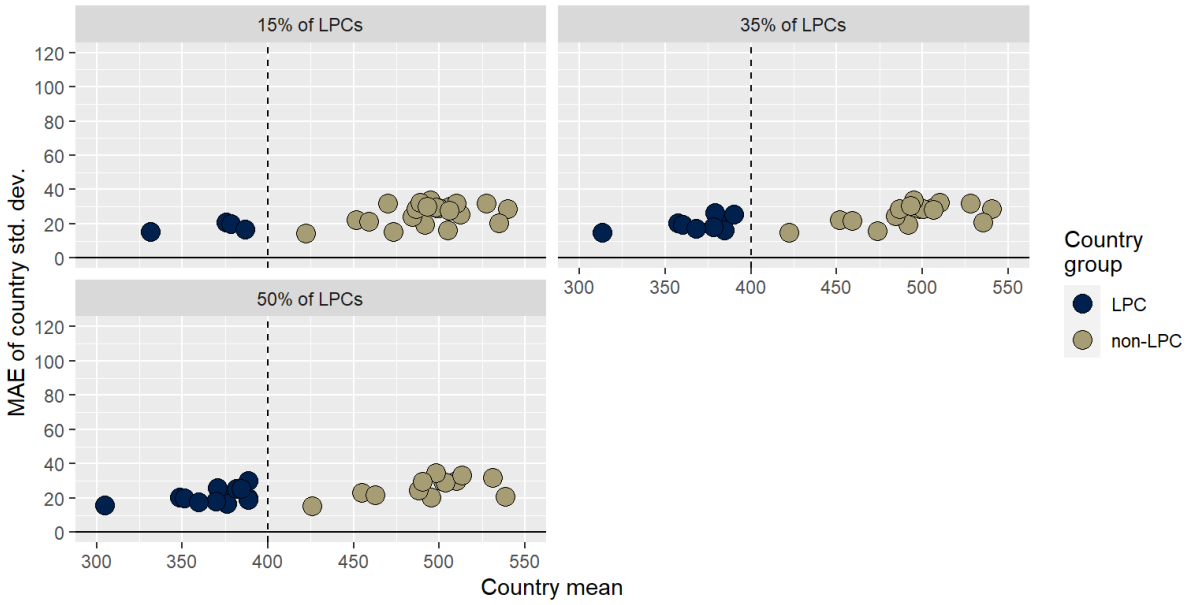
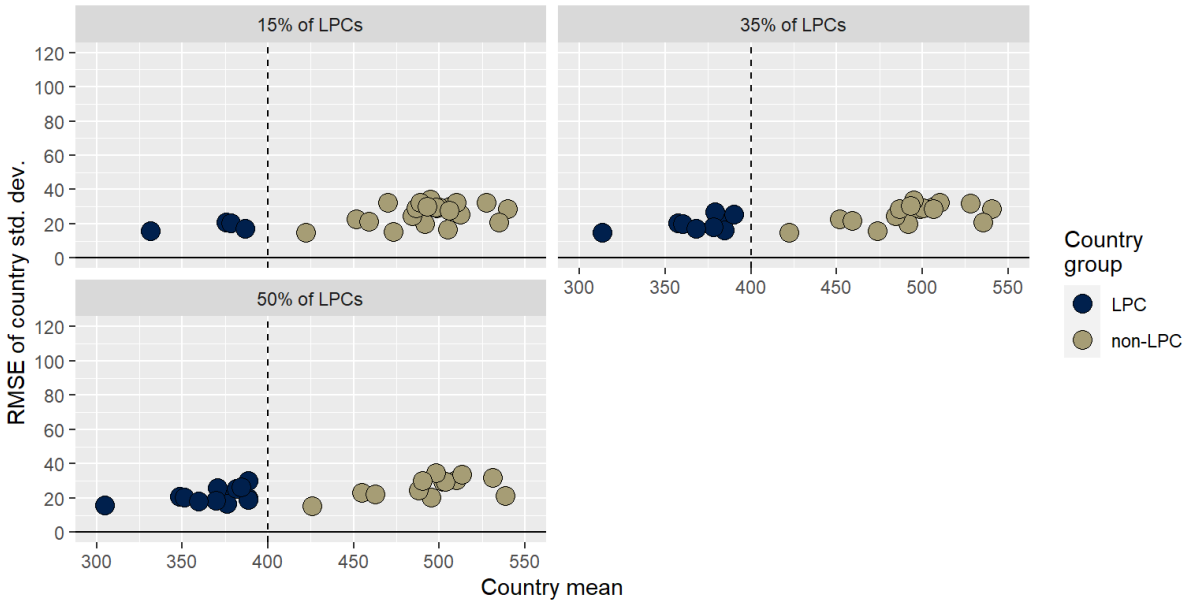


Figure A A.4. RMSE for country standard deviations for different proportion of LPCs and countries' mean proficiency



Additional graphs for 3.2

Figure A A.5. Recovery of item parameters for different proportion of LPCs in conditions with lower item difficulties of non-linking items

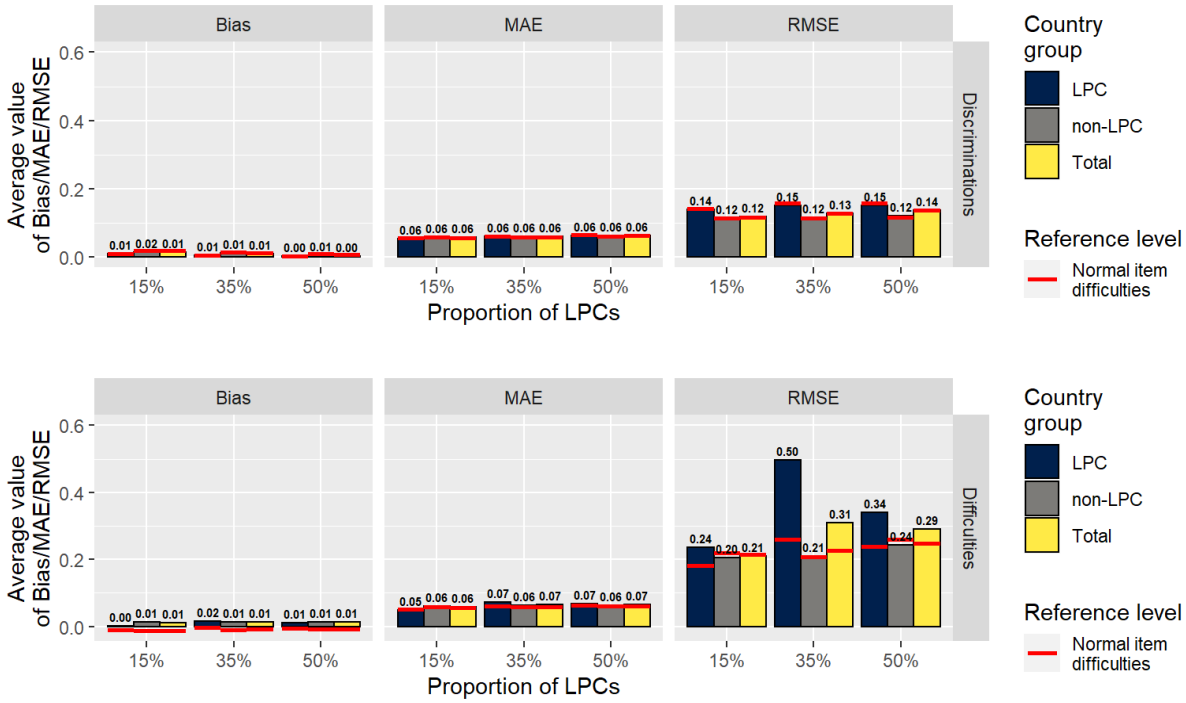


Figure A A.6. Recovery of item parameters for different proportion of LPCs in conditions with first calibration performed using only OECD countries

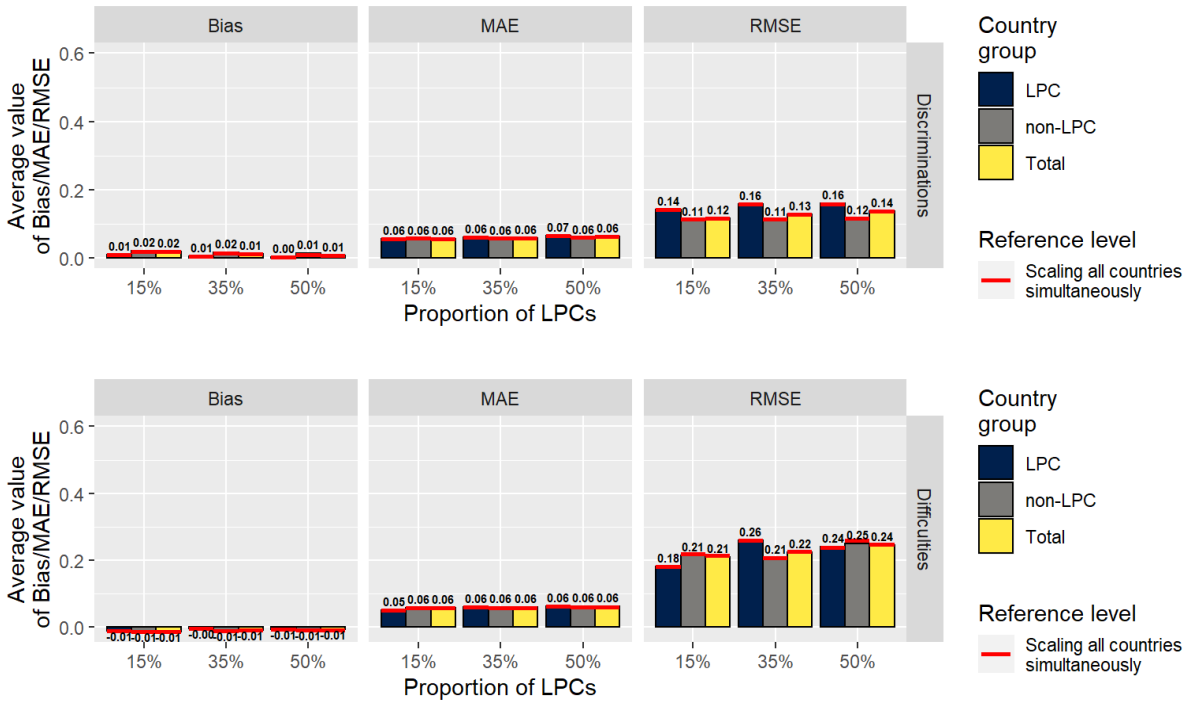


Figure A A.7. Recovery of item parameters for different proportion of LPCs in conditions with Perfect DIF detection

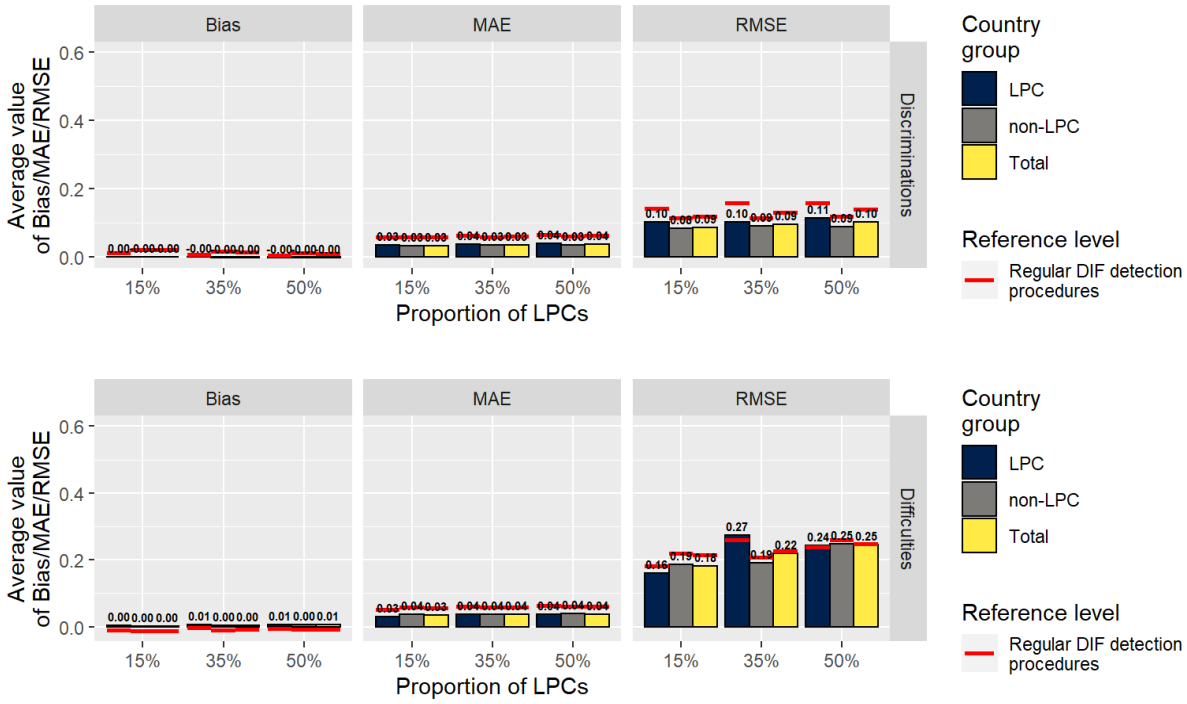


Figure A A.8. Recovery of item parameters for different proportion of LPCs in conditions with ignoring DIF (calibrating only the invariant model)

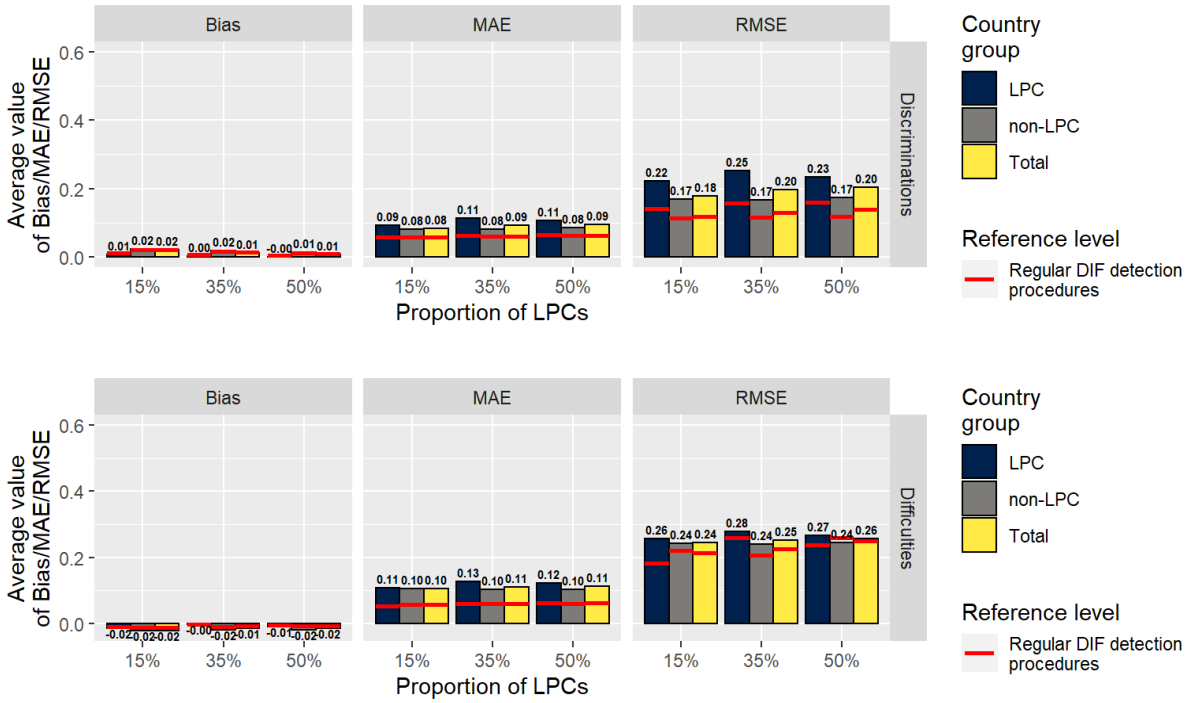


Figure A A.9. Recovery of DIF for different proportion of LPCs in conditions with lower item difficulties of non-linking items

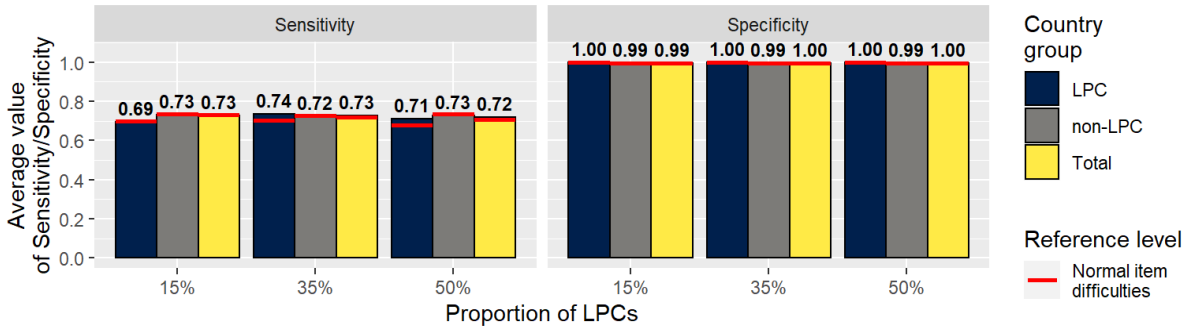
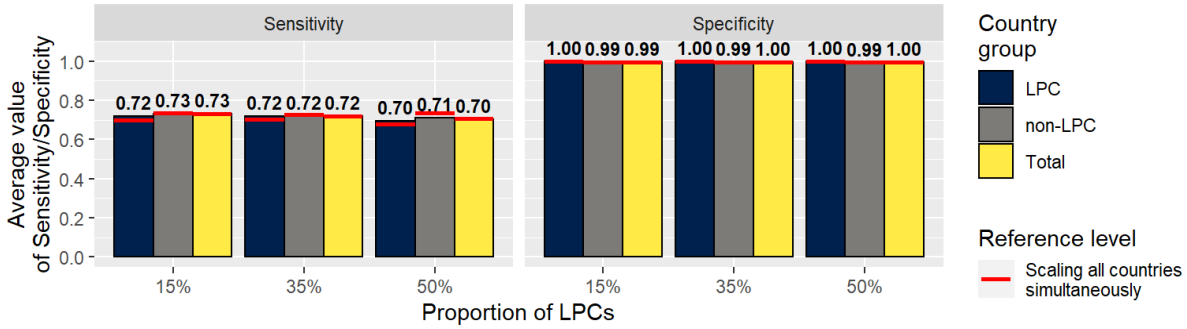


Figure A A.10. Recovery of DIF for different proportion of LPCs in conditions with first calibration performed using only OECD countries



Simulation code

```
conditionsFile <- "simulation_conditions.csv"
resultsFile <- sub("conditions", "results", conditionsFile)
repeatCondition <- 100
maxiter <- 500
RMSDthreshold <- c(0.3, 0.2, 0.12)
integrationGrid = seq(-4, 4, by = 0.2)
set.seed(25062111)

# setup ########################################################################

install.packages(c("remotes", "tidyr", "dplyr", "TAM"))
remotes::install_github("tzoltak/rstyles")

library(tidyr)
library(dplyr)
library(rstyles)
library(TAM)

load("PISA2015-data-to-simulation.RData")
# defining objects to store the results ########################################
conditions <- read.csv(conditionsFile, stringsAsFactors = FALSE) %>%
  expand_grid(k = 1:repeatCondition) %>%
  arrange(k)
resultsItems <- data.frame()
resultsGroups <- data.frame()
# function definitions #########################################################
prepare_fixed_pars_matrices <- function(model, fixedPars) {
  fixedXsi <- model$xsi.fixed.estimated %>%
        as.data.frame() %>%
        mutate(item = sub("_.*$", "", rownames(.)),
        variant = sub("^.*_([[:lower:]][[:digit:]]*)_.*$", "\\1", rownames(.)),
        variant = ifelse(rownames(.) == variant, "i", variant),
        category = sub("^.*Cat", "", rownames(.))) %>%
        select(index = V1, item, variant, category) %>%
        inner_join(fixedPars %>%
                select(item, variant, t1, t2) %>%
                pivot_longer(-c(item, variant),
                             names_to = "category", names_prefix = "t",
                             values_to = "xsi") %>%
                filter(!is.na(xsi))) %>%
        select(index, xsi) %>%
        as.matrix()
  fixedB <- model$B.fixed.estimated %>%
        as.data.frame() %>%
        setNames(c("itemIndex", "category", "dimension", "b")) %>%
        inner_join(model$item %>%
                select(item) %>%
                mutate(itemIndex = 1:n(),
                       variant = sub("^.*_", "", item),
                       item = sub("_.*$", "", item),
                       variant = ifelse(item == variant, "i", variant))) %>%
        inner_join(fixedPars %>%
                select(item, variant, a)) %>%
        mutate(a = ifelse(b != 0, a * (category - 1), 0)) %>%
        select(itemIndex, category, dimension, a) %>%
        as.matrix()
  return(list(Xsi = fixedXsi,
              B = fixedB))
}
get_estimated_item_pars <- function(model) {
  inner_join(
        model$xsi.fixed.estimated %>%
        as.data.frame() %>%
        mutate(item = sub("_.*$", "", rownames(.)),
                variant = sub("^.*_([[:lower:]][[:digit:]]*)_.*$", "\\1", rownames(.)),
                variant = ifelse(rownames(.) == variant, "i", variant),
                category = sub("^.*Cat", "", rownames(.))) %>%
        select(item, variant, category, xsi) %>%
        pivot_wider(names_from = category, names_prefix = "t", values_from = xsi),
        model$B.fixed.estimated %>%
```

```
        as.data.frame() %>%
        setNames(c("itemIndex", "category", "dimension", "a")) %>%
        filter(category == 2) %>%
        inner_join(model$item %>%
                       select(item) %>%
                       mutate(itemIndex = 1:n(),
                       variant = sub("^.*_", "", item),
                       item = sub("_.*$", "", item),
                       variant = ifelse(item == variant, "i", variant))) %>%
        select(item, variant, a)
  ) %>%
        return()
}
get_estimated_group_pars <- function(model) {
  cbind(GROUP = model$groups,
        mean = model$beta,
        cbind(GROUPno = model$group, sd = model$variance^0.5) %>%
        as.data.frame() %>%
        distinct() %>%
        arrange(GROUPno) %>%
        select(-GROUPno)) %>%
        return()
}
expand_items <- function(responses, itemsVariants) {
  responses %>%
        select(GROUP, all_of(unique(itemsVariants$item))) %>%
        mutate(id = 1:n()) %>%
        pivot_longer(-c(GROUP, id),
               names_to = "item", values_to = "score") %>%
        filter(!is.na(score)) %>%
        left_join(itemsVariants) %>%
        group_by(GROUP, item) %>% # check for no-variance item-groups
        mutate(variant = ifelse(n_distinct(score, na.rm = TRUE) < 2,
                                "i", variant)) %>%
        ungroup() %>%
        pivot_wider(names_from = c(item, variant), values_from = score) %>%
        select(-id) %>%
        return()
}
detect_items_variants_rmsd <- function(model, RMSDthreshold) {
  variants <- IRT.itemfit(model)$RMSD %>%
        select(-WRMSD) %>%
        pivot_longer(-item, names_to = "GROUPno", names_prefix = "Group",
               values_to = "RMSD") %>%
        filter(!is.na(RMSD))
  if (all(variants$RMSD <= RMSDthreshold)) {
        return(NULL)
  }
  variants %>%
        mutate(GROUPno = as.numeric(GROUPno),
        variant = sub("^.*_", "", item),
        item = sub("_.*$", "", item),
        variant = ifelse(item == variant, "i", variant),
        detectedVariant = sub("^x", "", variant) %>%
               ifelse(. == variant, "0", .) %>%
               as.numeric()) %>%
        left_join(data.frame(GROUP = model$groups,
                       GROUPno = 1:length(model$groups))) %>%
        group_by(item) %>%
        mutate(detectedVariant =
               max(detectedVariant) + cumsum(RMSD > RMSDthreshold),
        variant = ifelse(RMSD > RMSDthreshold,
                               paste0("x", detectedVariant),
                               variant)) %>%
        ungroup() %>%
        select(item, GROUP, variant) %>%
        return()
}
prepare_results_items <- function(itemsPars, data, suffix) {
  data %>%
        rename_with(~paste0(., "_i"), -c(GROUP, matches("_"))) %>%
        pivot_longer(-GROUP, names_to = c("item", "variant"), names_sep = "_",
               values_to = "score") %>%
```

```
        filter(!is.na(score)) %>%
        select(-score) %>%
        distinct() %>%
        inner_join(itemsPars %>%
                select(item, variant, a, t1, t2)) %>%
        rename_with(~ifelse(. %in% c("GROUP", "item"),
                        ., paste0(., suffix))) %>%
        return()
}
prepare_results_groups <- function(groupsPars, suffix) {
  groupsPars %>%
        rename_with(~ifelse(. %in% c("GROUP"), ., paste0(., suffix))) %>%
        return()
}
# simulation per se ###########################################################
for (i in 1:nrow(conditions)) {
  ## selecting the data #######################################################
  countriesI <- countries %>%
        filter(across(all_of(paste0("group_", conditions$prop_lpc[i])),
                ~. %in% c("LPC", "OECD")))
  itemsParsI <- itemsPars %>%
        filter(GROUP %in% countriesI$GROUP)
  conditions$items_pool_difficulty[i] <-
        match.arg(conditions$items_pool_difficulty[i], c("normal", "easier"))
  if (conditions$items_pool_difficulty[i] == "easier") {
        itemsParsI <- itemsParsI %>%
        select(-t1, -t2) %>%
        rename(t1 = t1Easier, t2 = t2Easier)
  }
  pisaI <- inner_join(countriesI %>%
                        select(GROUP, ADMINMODE),
                        pisa) %>%
        group_by(GROUP) %>%
        slice_sample(n = 5000, weight_by = SENWT, replace = TRUE) %>%
        ungroup()
  resultsItemsI <- itemsParsI %>%
        mutate(propLpc = conditions$prop_lpc[i],
        itemPoolDiffic = substr(conditions$items_pool_difficulty[i], 1, 1),
        k = conditions$k[i]) %>%
        select(propLpc, itemPoolDiffic, k, item, GROUP, variant, a, t1, t2)
  resultsGroupsI <- countriesI %>%
        mutate(propLpc = conditions$prop_lpc[i],
        itemPoolDiffic = substr(conditions$items_pool_difficulty[i], 1, 1),
        k = conditions$k[i]) %>%
        select(propLpc, itemPoolDiffic, k, GROUP, isLPC, ADMINMODE, mean, sd)
  ## generating values of the latent trait ###################################
  latentI <- countriesI %>%
        select(-nSchools, -starts_with("mean"),
        all_of(paste0("mean_", conditions$prop_lpc[i]))) %>%
        rename_with(~sub("^mean_.*$", "mean", .)) %>%
        mutate(sdBSchools = sd * (1 - icc)^0.5,
        sdWSchools = sd * icc^0,5) %>%
        left_join(pisaI %>%
                count(GROUP, CNTSCHID, name = "nStudents")) %>%
        group_by(GROUP, ADMINMODE, isLPC, sdWSchools) %>%
        summarise(schMean = rnorm(n(), mean, sdBSchools),
                nStudents = nStudents,
                .groups = "drop") %>%
        group_by(GROUP, ADMINMODE, isLPC, schMean, nStudents) %>%
        summarise(SCIE = rnorm(nStudents, schMean, sdWSchools),
                .groups = "drop") %>%
        select(GROUP, ADMINMODE, isLPC, SCIE) %>%
        nest(SCIE = SCIE)
  ## preparing list of test objects ##########################################
  testsI <- nest(itemsParsI, test = -GROUP) %>%
        mutate(test = lapply(test, function(x) {
        split(x, x$item) %>%
        lapply(function(x) {
        steps <- ifelse(is.na(x$t2), 1, 2)
        make_item(scoringMatrix =
                        matrix(0:2, ncol = 1, dimnames = list(0:2, "SCIE")
                        )[1:(steps + 1), , drop = FALSE],
                        slopes = setNames(x$a, "SCIE"),
```

```
                    intercepts = -as.vector(na.omit(unlist(x[c("t1", "t2")]))),
                    mode = "simultaneous") %>%
            return()
      }) %>%
      return()
      }))
## generating responses ###################################################
observedI <- inner_join(latentI, testsI) %>%
      group_by(GROUP, ADMINMODE, isLPC) %>%
      summarise(responses =
            list(cbind(SCIE = SCIE[[1]],
                        generate_test_responses(SCIE[[1]], test[[1]]) %>%
                        as.data.frame() %>%
                        setNames(names(test[[1]])))),
            .groups = "drop")
observedI <- inner_join(
      observedI %>%
      select(GROUP, ADMINMODE, isLPC),
      bind_rows(setNames(observedI$responses, observedI$GROUP),
            .id = "GROUP") %>%
      mutate(across(-c(GROUP, SCIE), as.numeric)))
## masking responses with respect to the real PISA data ######################
itemsNames <- unique(itemsParsI$item)
observedI <- observedI[, c("GROUP", "ADMINMODE", "isLPC", "SCIE", itemsNames)] %>%
      arrange(GROUP, SCIE)
pisaI <- pisaI[, c(setdiff(names(pisa), itemsNames), itemsNames)] %>%
      arrange(GROUP, PV1SCIE)
observedI[, itemsNames][is.na(pisaI[, itemsNames])] = NA
## scaling with perfect DIF detection #####################################
dataTemp <- expand_items(observedI,
                        itemsParsI %>%
                        select(item, GROUP, variant) %>%
                        distinct())
mPDD <- tam.mml.2pl(select(dataTemp, -GROUP),
                    group = dataTemp$GROUP,
                    irtmodel = "GPCM", control = list(maxiter = 1))
fixedPars <- prepare_fixed_pars_matrices(mPDD, linkingPars)
mPDD <- tam.mml.2pl(select(dataTemp, -GROUP),
                    group = dataTemp$GROUP,
                    irtmodel = "GPCM",
                    control = list(maxiter = maxiter, nodes = integrationGrid),
                    B.fixed = fixedPars$B, xsi.fixed = fixedPars$Xsi,
                    est.variance = TRUE)
mPDD <- mPDD[c("groups", "beta", "group", "variance",
            "xsi.fixed.estimated", "B.fixed.estimated", "item", "time")]
resultsItemsI <- resultsItemsI %>%
      left_join(prepare_results_items(get_estimated_item_pars(mPDD),
                                      dataTemp, "PDD"))
resultsGroupsI <- resultsGroupsI %>%
      left_join(prepare_results_groups(get_estimated_group_pars(mPDD), "PDD"))
## scaling - all countries at once ########################################
dataTemp <- observedI %>%
      select(GROUP, where(~is.numeric(.) & n_distinct(., na.rm = TRUE) > 1), -SCIE)
mA <- tam.mml.2pl(select(dataTemp, -GROUP),
                    group = dataTemp$GROUP,
                    irtmodel = "GPCM", control = list(maxiter = 1))
fixedPars <- prepare_fixed_pars_matrices(mA, linkingPars)
mA <- tam.mml.2pl(select(dataTemp, -GROUP),
                    group = dataTemp$GROUP,
                    irtmodel = "GPCM",
                    control = list(maxiter = maxiter, nodes = integrationGrid),
                    B.fixed = fixedPars$B, xsi.fixed = fixedPars$Xsi,
                    est.variance = TRUE)
itemsVariants <- detect_items_variants_rmsd(mA, RMSDthreshold[1])
parsMA <- get_estimated_item_pars(mA)
### saving results of the invariant model ################################
resultsItemsI <- resultsItemsI %>%
      left_join(prepare_results_items(get_estimated_item_pars(mA),
                                      dataTemp, "ID"))
resultsGroupsI <- resultsGroupsI %>%
      left_join(prepare_results_groups(get_estimated_group_pars(mA), "ID"))
### DIF detection procedure ##############################################
jA <- 1
```

```
t <- 1
while (!is.null(itemsVariants) & (jA <= 10 | t < length(RMSDthreshold))) {
      dataTemp <- expand_items(observedI, itemsVariants)
      mA <- tam.mml.2pl(select(dataTemp, -GROUP),
                        group = dataTemp$GROUP,
                        irtmodel = "GPCM", control = list(maxiter = 1))
      fixedPars <- prepare_fixed_pars_matrices(mA, parsMA)
      mA <- tam.mml.2pl(select(dataTemp, -GROUP),
                        group = dataTemp$GROUP,
                        irtmodel = "GPCM",
                        control = list(maxiter = maxiter, nodes = integrationGrid),
                        B.fixed = fixedPars$B, xsi.fixed = fixedPars$Xsi,
                        est.variance = TRUE)
      itemsVariants <- detect_items_variants_rmsd(mA, RMSDthreshold[t])
      if (t < length(RMSDthreshold)) {
      t <- t + 1
      itemsVariants <- detect_items_variants_rmsd(mA, RMSDthreshold[t])
      }
      parsMA <- get_estimated_item_pars(mA)
      jA <- jA + 1
}
resultsItemsI <- resultsItemsI %>%
      left_join(prepare_results_items(get_estimated_item_pars(mA),
                                      dataTemp, "All"))
resultsGroupsI <- resultsGroupsI %>%
      left_join(prepare_results_groups(get_estimated_group_pars(mA), "All"))
mA <- mA[c("groups", "beta", "group", "variance",
           "xsi.fixed.estimated", "B.fixed.estimated", "item", "time")]
## scaling - only OECD countries at first ###################################
### only OECD (also means only CBA)
dataTemp <- observedI %>%
      filter(!isLPC) %>%
      select(GROUP, where(~is.numeric(.) & n_distinct(., na.rm = TRUE) > 1), -SCIE)
mO <- tam.mml.2pl(select(dataTemp, -GROUP),
                  group = dataTemp$GROUP,
                  irtmodel = "GPCM", control = list(maxiter = 1))
fixedPars <- prepare_fixed_pars_matrices(mO, linkingPars)
mO <- tam.mml.2pl(select(dataTemp, -GROUP),
                  group = dataTemp$GROUP,
                  irtmodel = "GPCM",
                  control = list(maxiter = maxiter, nodes = integrationGrid),
                  B.fixed = fixedPars$B, xsi.fixed = fixedPars$Xsi,
                  est.variance = TRUE)
mO <- mO[c("xsi.fixed.estimated", "B.fixed.estimated", "item")]
### OECD & LPCs
dataTemp <- observedI %>%
      select(GROUP, where(~is.numeric(.) & n_distinct(., na.rm = TRUE) > 1), -SCIE)
mOL <- tam.mml.2pl(select(dataTemp, -GROUP),
                   group = dataTemp$GROUP,
                   irtmodel = "GPCM", control = list(maxiter = 1))
fixedPars <-
      prepare_fixed_pars_matrices(mOL,
                            bind_rows(get_estimated_item_pars(mO),
                                      linkingPars) %>%
                                  distinct())
mOL <- tam.mml.2pl(select(dataTemp, -GROUP),
                   group = dataTemp$GROUP,
                   irtmodel = "GPCM",
                   control = list(maxiter = maxiter, nodes = integrationGrid),
                   B.fixed = fixedPars$B, xsi.fixed = fixedPars$Xsi,
                   est.variance = TRUE)
itemsVariants <- detect_items_variants_rmsd(mOL, RMSDthreshold[1])
parsMOL <- get_estimated_item_pars(mOL)
### DIF detection procedure ################################################
jOL <- 1
t <- 1
while (!is.null(itemsVariants) & (jOL <= 10 | t < length(RMSDthreshold))) {
      dataTemp <- expand_items(observedI, itemsVariants)
      mOL <- tam.mml.2pl(select(dataTemp, -GROUP),
                         group = dataTemp$GROUP,
                         irtmodel = "GPCM", control = list(maxiter = 1))
      fixedPars <- prepare_fixed_pars_matrices(mOL, parsMOL)
      mOL <- tam.mml.2pl(select(dataTemp, -GROUP),
```

```
                        group = dataTemp$GROUP,
                        irtmodel = "GPCM",
                        control = list(maxiter = maxiter, nodes = integrationGrid),
                        B.fixed = fixedPars$B, xsi.fixed = fixedPars$Xsi,
                        est.variance = TRUE)
        itemsVariants <- detect_items_variants_rmsd(mOL, RMSDthreshold[t])
        if (t < length(RMSDthreshold)) {
        t <- t + 1
        itemsVariants <- detect_items_variants_rmsd(mOL, RMSDthreshold[t])
        }
        parsMOL <- get_estimated_item_pars(mOL)
        jOL <- jOL + 1
  }
  resultsItemsI <- resultsItemsI %>%
        left_join(prepare_results_items(get_estimated_item_pars(mOL),
                                        dataTemp, "OL"))
  resultsGroupsI <- resultsGroupsI %>%
        left_join(prepare_results_groups(get_estimated_group_pars(mOL), "OL"))
  mOL <- mOL[c("groups", "beta", "group", "variance",
               "xsi.fixed.estimated", "B.fixed.estimated", "item", "time")]
  # saving parameters ########################################################
  resultsItems <- bind_rows(resultsItems, resultsItemsI)
  resultsGroups <- bind_rows(resultsGroups, resultsGroupsI)
  write.csv(resultsItems, sub("\\.csv$", "_items.csv", resultsFile),
               row.names = FALSE, na = "")
  write.csv(resultsGroups, sub("\\.csv$", "_groups.csv", resultsFile),
               row.names = FALSE, na = "")
}
```