

Chapter 12

Differential Item Functioning in PISA Due to Mode Effects



Remco Feskens, Jean-Paul Fox and Robert Zwitser

Abstract One of the most important goals of the Programme for International Student Assessment (PISA) is assessing national changes in educational performance over time. These so-called trend results inform policy makers about the development of ability of 15-year-old students within a specific country. The validity of those trend results prescribes invariant test conditions. In the 2015 PISA survey, several alterations to the test administration were implemented, including a switch from paper-based assessments to computer-based assessments for most countries (OECD 2016a). This alteration of the assessment mode is examined by evaluating if the items used to assess trends are subject to differential item functioning across PISA surveys (2012 vs. 2015). Furthermore, the impact on the trend results due to the change in assessment mode of the Netherlands is assessed. The results show that the decrease reported for mathematics in the Netherlands is smaller when results are based upon a separate national calibration.

R. Feskens (✉)
Cito, Arnhem, The Netherlands
e-mail: remco.feskens@cito.nl

J.-P. Fox
University of Twente, Enschede, The Netherlands

R. Zwitser
University of Amsterdam, Amsterdam, The Netherlands

© The Author(s) 2019
B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_12

12.1 Introduction

The Programme for International Student Assessment (PISA) is an international comparative research project investigating the knowledge and skills of 15-year old students in science, reading literacy, and mathematics (OECD 2016a). Since 2000, PISA has been conducted every three years under the auspices of the Organization for Economic Cooperation and Development (OECD) with each administration featuring a different core component. The main aim of PISA is to provide participating countries with comparable information which can be used to evaluate and improve their educational policies. The evaluation of existing educational policy is mainly based upon the assessment of the functional literacy skills of students as measured by standardized cognitive tests. Country-specific performance can be compared to other countries within one administration, but country performances can also be compared between administrations. Comparisons based upon the latter inform countries if performance in one of the three subjects has changed.

The trend results for 2015 showed that many countries decreased in their overall performance when compared to the previous administration in 2012. This holds in particular for the core domain, science, but many countries also experienced a decrease in their mean mathematics performance. For example in the Netherlands, the mean country estimate for mathematics in 2015 decreased when compared to 2012, dropping 13 points (Feskens et al. 2016). This decrease is substantial, given that changes in country means measured over a short period of time are usually small (Mazzeo and von Davier 2008; cf. Robitzsch et al. 2017). This could imply that many of the 2015 country cohorts are less proficient when compared to their 2012 counterparts and that consequently, the overall educational level has dropped in these countries.

However, several changes were implemented in the 2015 PISA administration. The main changes in the 2015 methodology involved: (1) a change from traditional paper-based assessment (PBA) to a computer-based assessment (CBA) in most of the participating countries; (2) a different assessment of item nonresponse; (3) the use of a two-parameter logistic model (Birnbaum 1968; Muraki 1992) instead of a one-parameter logistic model (Rasch 1960; Masters 1982) for item response modelling; and (4) a change in the treatment of differential item functioning (DIF) across countries (OECD 2016b). The trend result decreases could also be due to these changes rather than just reflecting a decrease in academic performance.

Changing the mode of assessment administration can affect student performance and induce DIF (Kolen and Brennan 2014). Earlier studies have already discussed the validity of the trend results from PISA 2015. Robitzsch et al. (2017) concluded that an observed decrease in mathematics and science might also be due to unfamiliarity with computer use in classroom settings among German students. Robitzsch and colleagues (2017) came to this conclusion by re-analysing the PISA data and estimating the mean German performances comparing several calibration designs. Another recent study conducted by Jerrim et al. (2018) also questions the comparability of PBA and CBA results. Both studies are among others based upon a reanalysis of the

2014 PISA field trial study data available for some countries. This field trial served to ensure the comparability of test results across paper and computer modes (OECD 2016b). The OECD (2016b) concluded “there are very few samples where the main effect of the mode of delivery reaches statistical significance” (p. 15 Annex A6), but also recognizes that national field trial samples were small and not specifically designed to examine whether items administered in one mode were systematically easier or harder than those administered in another mode at the national level. Nevertheless, the random assignment of students within field trial schools made it possible to disentangle mode and population effects at the international level. This paper evaluates to what degree the new test administration mode of PISA affected items and the performance of the mathematics domain by comparing the last two main PISA cycles with a focus on the Dutch trend results. This will be done by first evaluating for DIF across modes and then by reanalysing the trend results for the Netherlands. We will re-estimate the Dutch results by only making use of the national data collected in the Netherlands. Consequently, only country-specific item parameters are included in this approach and the national trend estimates are not influenced by differences in mode effects between countries.

These considerations have led to following research questions:

1. To what extent are items used to establish trend effects in PISA 2015 affected by DIF?
2. To what extent do trend estimates for the Netherlands differ when results are only based upon the Dutch data, instead of the complete international data file?

This paper will continue with a description of the changes made in the 2015 PISA methodology and the data included for this study, followed by an overview of DIF and the impact DIF can have on trend results. Results will then be presented and followed by conclusions and discussion.

12.2 Changes in PISA 2015

Since its first administration in 2000, the number of countries and students participating in PISA has grown to a total of over 500,000 students from 72 different countries and economies that took part in 2015. All 35 OECD-member countries participated along with so-called partner countries without membership. All PISA research findings are presented using scales that are standardized to an international average of 500 with a standard deviation of 100. The average of 500 only applies to the OECD countries and was determined in the year in which a specific topic (reading literacy in 2000, mathematics in 2003, and science in 2006) dominated the agenda for the first time. The main mode of test administration in PISA was PBA up to 2012. This was changed in 2015, when most of the participating countries switched to CBA. The main reasons for this change are found in the creation of additional possibilities for measuring specific aspects of the domain constructs, the establishment of a more

rigorous procedure for test marking and the availability of detailed metadata, such as response time information (OECD 2017).

Despite these major advantages, there is the potential risk of incomparability of results between cycles, since student performances can be affected by aspects specifically related to the differences between CBA and PBA (Bridgeman et al. 2003; Leeson 2006). These differences have to do with the ease of reading texts, the ease of reviewing or changing answers to questions, the speed of test taking, the clarity of figures and diagrams, and the difference between responding on a computer and on an answer sheet (Kolen and Brennan 2014). Previous research on the impact of these two modes on mathematics testing performance has produced mixed results. Hong et al. (2007) did not find a main effect for the testing mode on the performance of K-12 student mathematics tests, whereas Kingston (2009) reported in a meta-analysis of 81 studies that there was a very small advantage for a paper administration of mathematics tests. Sandene et al. (2005) also reported a higher level of performance with PBA when compared to CBA in a study conducted by the National Assessment of Educational Progress Technology-Based Assessment Project. At the item level in that study, five percent more students responded correctly on paper than on a computer. Despite this mixed evidence, it was hypothesized that mathematics items used in PISA 2012 could be transposed onto a screen without affecting trend data in PISA 2015 by keeping the computer skills required to administer the test to a minimum (OECD 2016a).

A second change in the 2015 PISA methodology was the assessment of item nonresponse. First of all, non-reached items were not only treated as not administered in estimating item parameters, but also not taken into consideration in estimating person parameters in 2015. In previous cycles of PISA, only the former was true, whereas non-reached items were seen as incorrect responses while estimating person parameters (OECD 2016a). During the analyses of this study it turned out that a second change with respect to the treatment of item nonresponse was the classification of item nonresponse into different types of item missingness. This change will be discussed in more detail in the results.

The third change implemented in the PISA 2015 design was the item response theory (IRT) model used to calibrate the response data. While a one-parameter logistic model (1PL) was used from 2000 to 2012, a two-parameter logistic (2PL) IRT model was deployed in 2015. The 2PL model is more flexible, allowing different item discrimination parameters, but those parameters also introduce an empirically-based item contribution to ascertain ability. To minimize trend break effects, as many trend items as possible were estimated according to a 1PL model (OECD 2016a).

Finally, DIF was treated differently in PISA 2015. Up to 2012, only a small number of items were considered as “not administered” for some countries as a result of DIF, mainly due to translation or printing errors (OECD 2016a). In 2015, the PISA consortium took two different types of DIF into account: DIF due to mode effects and DIF due to country effects. Mode effects were taken into account based upon the results of the 2014 field trial and further scaling operations in the main study. Items were classified as either scalar, metric, or non-metric invariant items (Annex A5 OECD 2016c). In order to be able to meaningfully compare the latent

means and correlations across groups, both the factor loadings and intercepts should be the same across groups (scalar invariance) (Steenkamp and Baumgartner 1998; Meredith 1993), or in IRT terms: the difficulty and discrimination parameters should have the same value. Scalar invariant items were used for establishing trend effects (51 items in total), the metric invariant items (items where the discrimination parameters were equal across modes, but the items had different difficulty parameters) served to improve measurement precision (30 items), and items where no metric invariance could be established were not used in the 2015 CBA test. Furthermore, in 2015 (a limited number of) specific country-by-cycle item parameters were allowed in the case of country DIF (OECD 2016a). Both the 2012 and 2015 procedures aimed to come to a concurrent item parameter estimation using item responses from all participating countries and only introduced a country specific treatment in case of moderate to severe deviations from the overall item parameters. Within this paper, we will focus on the change in assessment mode and the effect mode differences (between countries) have on item parameters and ultimately the mean country estimates.

12.3 Data

To keep student response burden at an acceptable level while at the same time maintaining construct validity, PISA makes use of a multi-booklet design. Students administer different subsets of all questions included in the test, called test versions or booklets. This paper only included students who have administered one of the regular test versions. For Mathematics 2012, these were test versions 1–13 (OECD 2014). For 2015 PISA, the regular CBA test versions 43–54 (OECD 2017) were included.

The mean proficiency estimates of the 53 countries that participated in both PISA 2012 and in the CBA of PISA 2015 are displayed in Table 12.1, where the countries are ordered by the mean 2015 country ability estimates. The number of students within each country that completed one of the regular test versions are presented in the last two columns of the table. Note that these numbers are somewhat smaller compared the total number of students that participated.

As can be seen in Table 12.1, the majority of the 53 countries—especially those with above-average performance—are confronted with a decrease in mean mathematics performance, as expressed by Cohen's d (Cohen 1988). The effect size for the difference between the 2012 and 2015 mean estimates in the Netherlands is -0.12 . The numbers of students included in 2015 is much smaller compared to 2012, because mathematics was the core domain for 2012. In the years that a subject is the core domain, all students administer items that measure the core domain. Mostly, only one-third of the students within a country take a test of non-core domain subject in any given cycle.

Table 12.1 PISA participant country ability estimates and student participants

Country	Mean 2012	Mean 2015	Effect size	n 2012	n 2015
Singapore	573.47	564.19	-0.10	5522	2019
Hong Kong	561.24	547.93	-0.15	4509	1777
Macao	538.13	543.81	0.07	5320	1473
Chinese Taipei	559.83	542.32	-0.17	6033	2541
Japan	536.41	532.44	-0.05	6303	2185
Korea	553.77	524.11	-0.31	5031	1833
Switzerland	530.93	521.25	-0.11	11,200	2706
Estonia	520.55	519.53	-0.01	4760	1857
Canada	518.07	515.65	-0.03	21,352	6592
Netherlands	522.97	512.25	-0.12	4322	1743
Denmark	500.03	511.09	0.14	7351	2247
Finland	518.75	511.08	-0.10	8731	1930
Slovenia	501.13	509.92	0.10	5731	2072
Belgium	514.53	506.98	-0.08	8241	3108
Germany	513.52	505.97	-0.09	4834	2095
Poland	517.50	504.47	-0.15	4596	2064
Ireland	501.50	503.72	0.03	5002	2647
Norway	489.37	501.73	0.15	4622	1803
Austria	505.54	496.74	-0.10	4702	2283
New Zealand	499.75	495.22	-0.05	4285	1504
Russian Federation	482.17	494.06	0.15	5207	2015
Sweden	478.26	493.92	0.18	4669	1781
Australia	504.15	493.90	-0.11	14,348	4783
France	494.99	492.92	-0.02	4542	2013
United Kingdom	493.93	492.48	-0.02	12,632	4670
Czech Republic	498.96	492.32	-0.07	5224	2249
Portugal	487.06	491.63	0.05	5651	2398
Italy	485.32	489.73	0.05	30,948	3830
Iceland	492.80	488.03	-0.05	3500	1117
Spain	484.32	485.84	0.02	25,189	2225
Luxembourg	489.85	485.77	-0.05	5246	1721
Latvia	490.57	482.31	-0.11	4306	1591
Lithuania	478.82	478.38	-0.01	4616	2157

(continued)

Table 12.1 (continued)

Country	Mean 2012	Mean 2015	Effect size	n 2012	n 2015
Hungary	477.04	476.83	0.00	4774	1860
Slovak Republic	481.64	475.23	-0.07	4607	2066
Israel	466.48	469.67	0.03	4993	2258
United States	481.37	469.63	-0.14	4947	1873
Croatia	471.13	464.04	-0.08	5003	1919
Greece	452.97	453.63	0.01	5115	1820
Bulgaria	438.74	441.19	0.03	2424	1958
United Arab Emirates	434.01	427.48	-0.07	5246	4629
Chile	422.63	422.67	0.00	3115	2315
Turkey	447.98	420.45	-0.34	4839	1924
Uruguay	409.29	417.99	0.11	2448	1986
Montenegro	409.63	417.93	0.11	4712	1875
Thailand	426.74	415.46	-0.15	6602	2719
Mexico	413.28	408.02	-0.08	15,398	2522
Qatar	376.45	402.40	0.28	10,831	5518
Costa Rica	407.00	400.25	-0.11	2028	2036
Colombia	376.49	389.64	0.19	3643	3872
Peru	368.10	386.56	0.24	2767	2309
Brazil	388.51	377.07	-0.15	8796	7614
Tunisia	387.82	366.82	-0.28	1990	1700

51 items were administered in both 2012 and 2015 PISA,¹ 31 of which have been classified as scalar invariant items that serve as anchor items. Ultimately, changes in overall mean country performances across PISA cycles is based upon student achievement in these items and, as such, they will be the subject of a DIF study in the following section.

12.4 Differential Item Functioning

Concerns related to how the test administration mode might affect the performance of students is the domain of measurement invariance. The measurement instrument—the cognitive test—should function in the same way across varied conditions, as long as these conditions (here the administration mode) are irrelevant to the attribute being measured (Millsap 2011). As soon as comparisons between in this

¹In PISA 2012, the item labels included an additional “P”, indicating they were administered on paper. The 2015 CBA item labels start with an additional “C”.

case two different cohorts are made, it is expected that the two tests will not produce systematically different results for groups with the same ability level. If consistent differences in the probability of endorsing an item are found for groups with the same ability level, the item is said to exhibit measurement bias with respect to the mode of administration (Millsap 2011). The evaluation of measurement invariance concerns the assessment how an item functions within different groups.

Therefore, item bias is also referred to by a more neutral term, DIF (Mellenbergh 2011). DIF is directly related to the evaluation of item response probabilities for members of different groups after matching on the latent trait that the test is intended to measure. DIF is a term coined for situations containing a reference group, defined here as the group of students who completed the PISA test using paper and pen, and a focal group, defined as the students who completed the test on a computer. DIF exists when the performance on a particular item or class of items differs between the focal and reference group, controlling for the performance on the test as a whole.

Although the concept of DIF seems straightforward, some problems have been highlighted in among others a recent study by Bechger and Maris (2014) and are mostly related to comparing parameters that are not identified from the observations. Bechger and Maris (2014) proposed using a differential item pair functioning DIF test, which focuses on comparing item pairs instead of seeing DIF as an item property. The difference with traditional procedures is that DIF is defined in terms of the relative difficulties of pairs of items—which are identified from the observations—and not in terms of the difficulties of individual items.

The procedure starts with a separate calibration of the data within each group. There exists an overall test for DIF, which under the null hypothesis that there is no DIF follows a Chi-square distribution with the number of items minus one degrees of freedom (Bechger and Maris 2014). If an item pair in the calibration of one group has a different relative difficulty when compared to the relative difficulty in the calibration of the second group, that item pair is subject to DIF. Differences between item pair difficulties can be tested using a Wald test and the results are usually summarized in a heat map. The plot highlights item pairs that have large inter-group differences in the relative positions of their relative difficulties (Bechger and Maris 2014).

Trends within countries can be estimated by making use of data collected in every country. This is done by making use of concurrent item parameter estimation aimed to place all items on a common international scale. Until PISA 2012 this was established in two stages: First, based upon a separate calibration in each country, items with poor psychometric properties were removed. Only in the second stage, a common scale for all countries was assumed (OECD 2009). In PISA 2015, a concurrent calibration was directly applied. Mainly based upon the field trial results, items administered in both assessment modes were constrained to have the same slope and threshold parameters (scalar invariant), or only the slope parameter was constrained to be the same across modes (metric invariant). Above that, in a relatively small number of cases, item constraints were released to allow the estimation of unique (national) item parameters (PISA 2017). Country trends can, however, also be established by estimating item parameters that only make use of national data, which has similarities to the DIF approach as proposed by Zwitser et al. (2017). This latter approach has the

advantage that country specific mode effects are taken into account more explicitly by allowing DIF between countries. For the purpose of estimating country trends based on national data only, this paper will calibrate item responses on the mathematics items collected in 2012 and 2015 from the Netherlands by employing an extended nominal response model (ENORM) using conditional maximum likelihood (CML)² and the same classifications of scalar and metric invariant items as used in PISA 2015. Person parameters will be estimated by drawing five plausible values for each student.

The results section will evaluate if the mathematics items measured in 2012 PISA by PBA and 2015 PISA by CBA are subject to DIF. Trend estimates will be compared based upon a separate analysis carried out on the Dutch national data. All analyses were conducted in R (R Core Team 2016) using the package “dexter” (Maris et al. 2018). In order to take the data collection design of PISA into account, the R package “survey” (Lumley 2010) was used to estimate the means for the Netherlands in 2012 and 2015.

12.5 Results

PISA 2012 made a distinction between four different types of missing data for the cognitive tests: (1) Item level nonresponse, which indicated an answer to a question was expected, but no response was given by the student; (2) Multiple or invalid responses, referring to (among others) instances of a student selecting multiple answer categories when only one response was expected; (3) Missing by design, referring to questions that were not included in the test version that the student administered or items that were deleted after the assessment because of misprints or translation errors; and (4) Non-reached items, covering consecutive missing values clustered at the end of a test session, except for those coded as item level nonresponse (OECD 2014).

In 2015, five different missing data types were used and classifications were changed. Item level nonresponse was termed “No response/omit” in 2015 and multiple or invalid responses were labelled simply “invalid”. The missing by design category was combined with missingness due to students ending the assessment early. Non-reached items remained untouched but the fifth category, “Not Applicable”, noted responses for questions that the respondent was directed to skip and responses that could not be determined due to printing problems or torn booklets (OECD 2017).

Figure 12.1 displays the item response category percentages for each item administered in 2012 and 2015. For comparison’s sake, the 2015 invalid and not applicable

²The IRT model used in PISA 2012 is the Mixed Coefficients Multinomial Logit Model using marginal maximum likelihood (MML) estimation (Adams et al. 1997). For dichotomous items and polytomous items both the ENORM and the Mixed Coefficients Multinomial Logit Model default to the Rasch or the partial credit model respectively (cf. <https://dexterities.netlify.com/2018/08/21/dexter-meets-pisa-1/>).

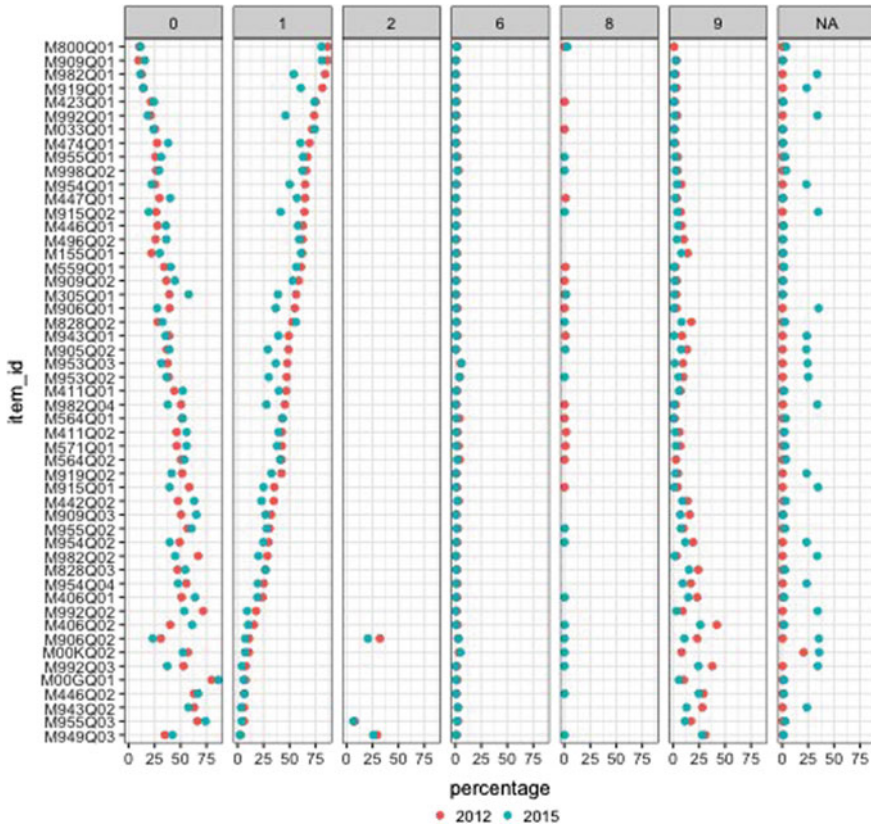


Fig. 12.1 Response category percentages PISA 2012 and 2015

item types were collapsed and compared to the 2012 “Multiple or invalid responses” category.

The first three panels in Fig. 12.1 show respective responses coded as zero (no credit), one (partial credit for polytomously scored items or full credit for dichotomously scored items), and two (full credit). Category six denotes non-reached items, category eight shows not applicable/invalid answers, and category nine corresponds to item nonresponse. Category NA shows that the missing by design (NA) category is also observed within test versions.³ As mentioned, within the 2012 administration this applies to items that were deleted after assessment because of misprints or translation errors. In 2015 these values within test versions apply to students who did not see the question due to an early ending of the assessment. While that last category had small percentages within the 2012 administration, its values became substantial in 2015 and, for some of the items, up to 30% of the students ended their test earlier than expected. Item nonresponse was slightly higher in 2012 while the response per-

³http://www.oecd.org/pisa/data/2015database/Codebook_CMB.xlsx.

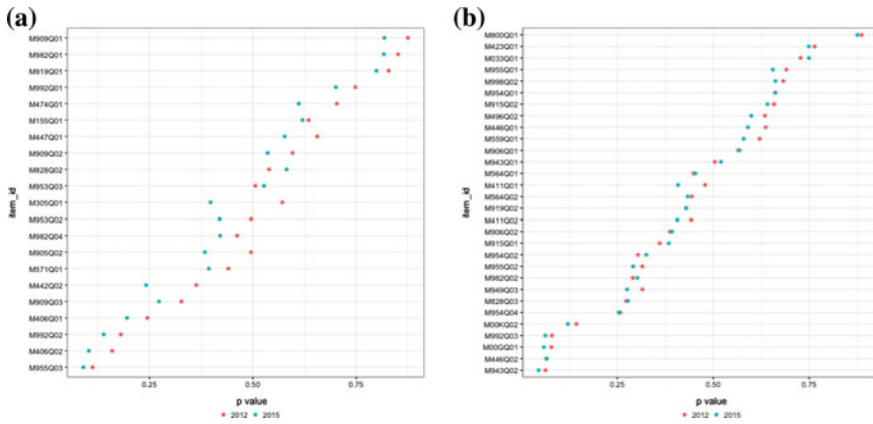


Fig. 12.2 P-values of metric (a) and scalar (b) invariant items for selected countries in 2012 and 2015

centages for the other missing data categories (non-reached and invalid) were similar across 2012 and 2015. It is important to note that responses coded as non-reached and premature ending of the assessment are considered as “not administered” in the computation of item difficulties and the IRT scaling of items. Figure 12.2 displays the item difficulties, or P-values, of the common items for the selected countries within the included test versions. Items classified by the OECD as metric invariant are found in the left panel and scalar invariant items are presented in the right panel.

As expected, the item difficulty rates of metric invariant items differ substantially between PBA and CBA. As aforementioned item difficulty parameters are also estimated separately for these items. The item difficulty rates of the scalar invariant items are comparable, though the values are somewhat higher in 2012. Nevertheless, the results suggest that the classification into scalar and metric invariant items based upon the field trial results has been advantageous in order to take mode effects into account.

12.5.1 DIF Between Modes

The overall test for DIF between modes indicates that even items classified as scalar invariant items might be subject to DIF (Chi-square = 2209, $df = 31$, $p < .01$). This indicates that although the procedure used by the consortium to take DIF into account was probably beneficial, there is still reason to believe that results could be subject to DIF. Figure 12.3 provides a visual summary of the item pair DIF Wald test results.

The results found in Fig. 12.3 suggest that especially item M033Q01 shows some indications for DIF. This is actually an item with a higher p-value in 2015. Although

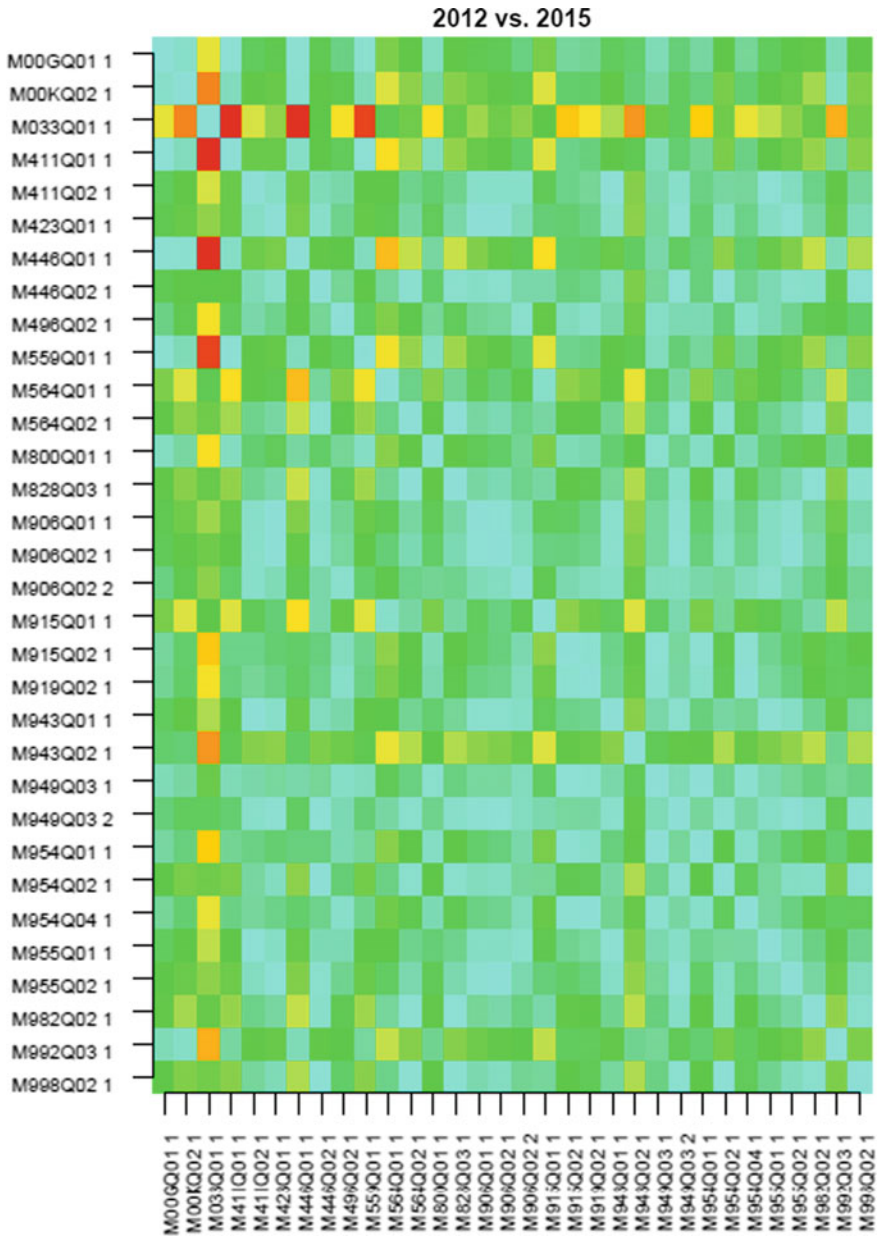


Fig. 12.3 Heat map presenting item pair DIF test results

the relative difficulty of many items in 2015 remain largely unchanged compared to 2012, both the overall test for DIF and the Wald test results for differences between item pair difficulties suggest that results might still be subject to DIF.

12.5.2 Trend Effects in the Netherlands

Figure 12.4 displays the P-values of the trend items in 2012 and 2015 based solely upon the PISA data collected in the Netherlands.

No clear trend can be detected in the unweighted 2012 and 2015 difficulty rates. To assess the mean trend effect using an alternative IRT scaling procedure, only data collected from the Netherlands in 2012 and 2015 were used to estimate item parameters. Based upon these country-specific item parameters the Dutch mean country estimates in 2012 and 2015 have been re-estimated. By applying this approach, the

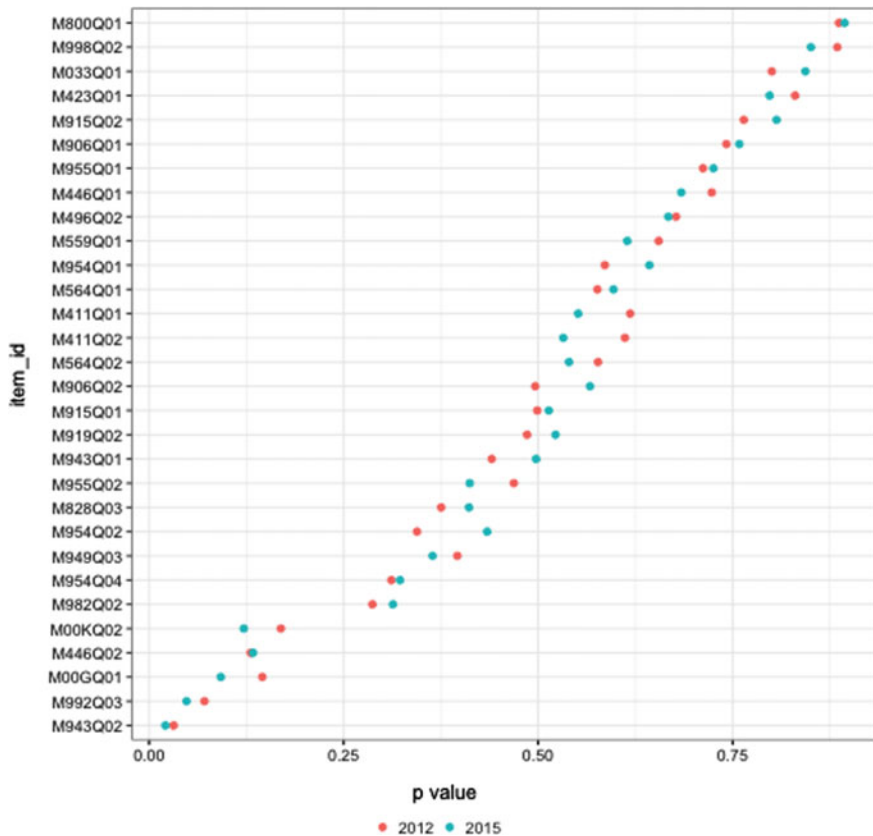


Fig. 12.4 P-values of trend items in the Netherlands for 2012 and 2015

weighted effect size for the differences in mean estimates in the Netherlands is -0.05 , about half the size of the reported effect size of -0.12 .

12.6 Conclusions and Discussion

Several key changes have been implemented in PISA 2015 and probably the most prominent one was the change of the test administration mode from PBA to CBA. This was a necessary step from many perspectives. The main potential advantage of this move is that a digital test platform facilitates a better representation of the PISA measurement objectives, which emphasize functional knowledge and skills needed to participate in society (OECD 2016a). In particular, the mathematical skills required at work merge with computer use (Hoyles et al. 2002; OECD 2016a). A well-known risk of changing the administration mode is that doing so might jeopardize the comparability of survey results. The contractors of PISA 2015 have taken many measures to accommodate potential trend breaks caused by the switch in the testing mode. A field trial was organized in 2014, prior to the main cycle, to test for mode effects and evaluate the comparability of test results across paper and computer modes (OECD 2016b). Based upon the field trial results, items were classified into scalar, metric and non-metric invariant items with only the first one being used to compare results across modes (OECD 2016c). However, reanalyses of the field trial data collected in Germany found national level mode effects where items administered on a computer were more difficult compared to a paper and pen administration (Robitzsch et al. 2017). Jerrim et al. (2018) concluded (based on a reanalyses of field trial data from Germany, Sweden, and Ireland) that the measures taken during the 2015 cycle have reduced, but not completely resolved, the impact of mode effects.

This study has assessed the extent to which mathematics items, measured in PISA 2012 and 2015 from countries participated using PBA in 2012 and CBA in 2015, are subject to DIF. The performance on scalar invariant items and the evaluation of DIF comparing the main survey results for both years demonstrates that the methodology PISA has adopted was beneficial in accounting for the mode effects. However, the reported decrease for mathematics in the Netherlands between 2012 and 2015, an effect size of -0.12 , could not be reproduced with a separate national scaling. Still, a decline in scores was found, but now with a smaller effect size of -0.05 . Thus, once DIF between countries is explicitly allowed, the decrease in trend results for the Netherlands is not as large as the reported decrease based upon the calibration procedure applied in PISA 2015. Furthermore, an increase in the number of students ending the assessment early or failing to see a given question was noted.

The reported decrease in mean science scores in 2015 was larger than the decrease in mathematics.⁴ This might be due to the introduction of new interactive items, which only took place for the science domain. The decrease in science might be relevant for the scores in mathematics as well, as the majority of reported mathematics plausible values are among other based upon the science scores. The results of this study are only based on the results of students that have been administered one of tests measuring mathematics, which ensures results are not confounded by the potential mode effects of other domains but limits the results to being based upon a subsample of students.

Although many publications on PISA have been made available by the OECD, some information on mode effects and how these effects have been taken into account is missing. For example at the time of writing of this article, the international data from the 2014 field trial were not publicly available. This makes it difficult to fully replicate the analyses carried out by the OECD. Nevertheless, given the results found in this and other recently published studies, the trend outcomes reported for the 2015 PISA should be interpreted with care.

References

- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.
- Bechger, T. M., & Maris, G. (2014). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340. <https://doi.org/10.1007/s11336-014-9408-y>.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley. <https://doi.org/10.2307/2283550>.
- Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*, 191–205. <https://doi.org/10.1002/j.2333-8504.2001.tb01865.x>.
- Cohen J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates. <https://doi.org/10.2307/2290095>.
- Feskens, R., Kuhlmeier, H., & Limpens, G. (2016). *Resultaten pisa-2015. Praktische kennis en vaardigheden van 15-jarigen*. Arnhem: Cito. URL: http://www.cito.nl/onderzoekenwetenschap/deelname_int_onderzoek/pisa.
- Hong, J., Young, M. J., Brooks, T., Olson, J., & Wang, S. (2007). A meta-analysis of testing mode effects in grade k-12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219–238. <https://doi.org/10.1177/0013164406288166>.
- Hoyles, C., Wolf, A., Molyneux-Hodgson, S., & Kent, P. (2002). *Mathematical skills in the workplace: Final report to the science technology and mathematics council*. Technical report, Institute of Education, University of London. URL <http://discovery.ucl.ac.uk/10001565/1/Hoyles2002MathematicalSkills.pdf>.

⁴The average of the mean science performance among the 35 OECD countries in 2012 was 501 and decreased to 493 in PISA 2015. For mathematics, this decrease was somewhat less profound: 494 in PISA 2012 to 490 in 2015 (OECD 2016b).

- Jerrim, J., Micklewright, J., Heine, J., Salzer, C., & McKeown, C. (2018). Pisa 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, 44(4), 476–493. <https://doi.org/10.1080/03054985.2018.1430025>.
- Kingston, N. (2009). Comparability of computer-and paper-administered multiple-choice tests for k-12 populations: A synthesis. *Applied Measurement in Education*, 22, 22–37.
- Kolen, M. J., Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York: Springer. <https://doi.org/10.1007/978-1-4939-0317-7>.
- Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, 6(1), 1–24. https://doi.org/10.1207/s15327574ijt0601_1.
- Lumley, T. (2010). *Complex surveys*. Wiley-Blackwell. <http://dx.doi.org/10.1002/9780470580066>.
- Maris, G., Bechger, T., Koops, J., Parchev, I. (2018) *dexter: Data management and analysis of tests*. URL <https://CRAN.R-project.org/package=dexter>. R package version 0.8.1.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/bf02296272>.
- Mazzeo, J., & von Davier, M. (2008). *Review of the programme for international student assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Technical report, Education Working Papers.
- Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics. Development, analysis, and application of psychological and educational tests*. The Hague, Netherlands: Eleven international publishing.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/bf02294825>.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. London: Routledge. <https://doi.org/10.4324/9780203821961>.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600206>.
- OECD. (2009). *PISA 2006 technical report*.
- OECD. (2014). *Pisa 2012 technical report*. Technical report, OECD, Paris.
- OECD. (2016a). *Pisa 2015 assessment and analytical framework: Science, reading, mathematics and financial literacy*. Technical report, OECD, Paris. URL <http://dx.doi.org/10.1787/9789264255425-en>.
- OECD. (2016b). *Pisa 2015 results (volume i): Excellence and equity in education*. Technical report, OECD Publishing, Paris.
- OECD. (2016c). *Pisa 2015 results (volume ii): Policies and practices for successful schools*. Technical report, OECD Publishing, Paris.
- OECD. (2017). *Pisa 2015 technical report*. Technical report, OECD Publishing, Paris. URL: <http://www.oecd.org/pisa/data/2015-technical-report/>.
- R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL www.R-project.org/.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Vol. 77). Danisch Institute for Educational Research. 10. 2307/2287805.
- Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica*, 63(2), 148–165. <https://doi.org/10.1026/0012-1924/a000177>.
- Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the naep technology-based assessment project*. Technical report, US Department of Education, National Center for Education Statistics.
- Steenkamp, J. M., Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–90, 25(1), 78–107. <https://doi.org/10.1086/209528>.

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at dif in international surveys. *Psychometrika*, 82(1), 210–232. <https://doi.org/10.1007/s11336-016-9543-8>. ISSN 1860-0980.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

