PISA Trends within the Netherlands

Evaluating mode effects

Stichting Cito

Onderzoek, Kennis & Innovatie

Auteurs
Konrad Klotzke
Remco Feskens

1. Executive Summary

Within this study we have evaluated the effect of transitioning from paper-based to computer-based assessments on the national PISA trends within the Netherlands. The PISA scores are designed to allow for comparisons between countries and across time. When neglecting the comparability across countries, the country-specific trends can be estimated with a higher accuracy. In this study we explore if items are appropriate to measure the national trends in the three PISA main domains within the Netherlands. The Netherlands-specific analyses are further refined to detect and control for mode and cycle effects that vary between genders or educational programs.

A different factor that is suspected to distort the international PISA rankings and trends is the varying effort that students invest into low-stake assessments. In an additional analysis, we explore the relationship between self-perceived effort and student performance across countries and within the Netherlands.

Adjusted national PISA trends

The estimated reading literacy within the Netherlands saw an upward correction at the 2018 PISA cycle when the trend results are based on a country-specific trends estimation procedure. The estimates at the 2015 and 2018 cycles were not affected by the country-specific trends analysis. It is unclear whether the shift in reading performance at the 2018 cycle can be attributed to Netherlands-specific differential item functioning (DIF) in anchor items, to an effect of the the in 2018 introduced multi-stage test design on student variables (e.g., the level of motivation or commitment during the test-taking process), or to a mixture of both.

The results indicate that for the mathematics domain, a global analysis or calibration is sufficient to capture possible item-mode and item-cycle interactions within the Netherlands. It was therefore not evident that the DIF detected in the analysis of data from the Netherlands reflects a systematic country-specific effect on the functioning of items. Instead, the

DIF could be attributed to random variation in the data collection process.

For the 2015 PISA cycle, the global calibration of anchor items led to an overestimation of the scientific literacy within the Netherlands. The adjusted trends show a strong decline in ability between 2012 and 2015, and an equal level of performance across the computer-based cycles. The results are indicative of a mode effect within the Netherlands that explains the drop in science performance between the paper- and computer-based cycles. The suspected mode effect did not systematically differ between female and male students but showed an interaction with the followed educational program.

The effect of motivation

On country-level, student effort correlated positively with growth in performance between the 2015 and 2018 PISA cycles. The correlation was weak for the reading domain, and of moderate size for the mathematics and science domains. In the international comparison of reading performance at the 2018 cycle, students in the Netherlands on average performed worse than expected given their reported level of invested effort. For the mathematics and science domains, the performance of the Netherlands at the 2018 cycle did not noticeably differ from the performance of other countries that reported a similar level of student effort.

The self-reported effort of students in the Netherlands was moderately positively correlated with their estimated ability at the 2018 PISA cycle. The correlation was higher for female than male students and noticeably weaker for students in the VWO program. The findings were consistent across the reading, mathematics and science domains.

Contents

2. Introduction

Every three years the OECD Programme for International Student Assessment (PISA) measures the proficiency of 15-year-old students worldwide in reading, mathematics and science literacy (OECD, 2019a). The PISA assessment is designed to produce measurements that are on the same scale across and within countries. Thereby countries can be ranked by their performance and the change in performance between PISA cycles can be monitored.

With the transition from paper-based to computer-based assessments the requirement for comparability of measurements extends to test modes. The term mode effect describes the case where an item functions systematically different for equivalent populations of students depending on the presentation mode of the assessment. Detecting, and if necessary controlling for, possible mode effects is therefore crucial to ensure that results from the paper- and computer-based assessment variants can be reported on a common scale. Mode effects that are overlooked can systematically favour, or disadvantage, students in one country or cycle and can thereby distort the rankings between, and the measured trends within countries.

The effect of presentation mode on item functioning was investigated during the field trial for the 2015 PISA cycle (OECD, 2016). Students from 68 countries were presented either the paper- or the computer-based variant of a shortened version of the PISA assessment. Statistical methods based on Item Response Theory (IRT) were applied to the data to identify trend items that showed a mode effect. For the following main survey part of the 2015 PISA cycle, a statistical model that controls for the detected mode effects was constructed. The appropriateness of the statistical model was evaluated for all country-by-language-by-cycle groups through group-specific item-fit statistics. To prevent a biased comparison between groups, the identified group-specific mode effects were accounted for in the statistical model.

The scaling procedure implemented in the 2015 cycle is designed to maximize the comparability across countries and is less suited to detect mode effects in individual countries. The limitations are evident by three key elements of the procedure. First of all, a possible student-item-mode interaction was evaluated through a concurrent analysis of the entire field trial sample. Based on model evidence, it was concluded that mode effects differed between items, but not students. However, given the moderate number of countries included in the field trial, this approach is insensitive to mode effects that occur in a single country or in a small subset of countries. The issue is aggravated by the heterogeneity within countries, i.e., a group-specific mode effect can be limited to certain subpopulations within a country (e.g., male students), which further decreases the statistical power to detect the country's deviation from the averaged, country-unspecific mode effect and can thereby under- or overestimate the country's performance in the international rankings or relative to other PISA cycles.

Secondly, a further limitation of the statistical model with which the student-item-mode interaction was investigated is the assumption of independence between the student-specific mode effects and their ability level. While it can be argued that a student's domain-specific proficiency is unlikely to directly impact their capability of processing an item and entering a response on a computer, it is plausible that constructs such as motivation or intelligence serve as confounding variables. In that case, constraining the correlation between ability and mode effect a priori to zero poses a model misspecification that reduces the statistical evidence for a student-item-mode interaction and thus lowers the probability of detecting country-specific mode effects.

Thirdly, cut-off based item-fit statistics were utilized to identify additional mode effects in the country-by-language-by-cycle groups from the main survey data. This technique flags an item as functioning differently in a group if the corresponding fit statistic passes a certain, pre-defined threshold. The concept seems akin to the level of significance that is applied in statistical hypothesis testing, however cut-off values lack a sound mathematical foundation (Marsh et al., 2004). Instead, the cut-off values are rules-of-thumb that are derived from former studies of similar data and populations, and ideally undergo further evaluation and fine-tuning in extensive simulations (Putnick & Bornstein, 2016). The heterogeneity of the PISA sample poses an additional difficulty in establishing appropriate cut-off values. In their study of the 2015 PISA data, Tijmstra et al. found that the fit statistics behaved differently for low- and high-performing groups, which could have exaggerated the differences in the country rankings and in the national trends. They moreover concluded that even under ideal conditions (correct and incorrect responses to an item are equally likely within a group) the cut-off value applied in the PISA study could often have left group-specific item functioning of moderate effect size undetected.

When neglecting the comparability across countries, country-specific analyses of the PISA data can be conducted that allow for a higher level of statistical rigor in detecting item-mode and item-cycle interactions. The analyses can be further refined to detect and control for mode and cycle effects in subpopulations of a country (e.g., genders or educational programs). Thereby new national trends can be estimated that reflect the development of performance within a country more accurately. Furthermore, the results from the 2018 PISA study were linked to previous cycles through the item parameters established in the 2015 cycle (OECD, 2019b). As a consequence, mode effects that went undetected in the 2015 scaling procedure carried over to the next cycle, which further motivates conducting country-specific analyses if the national trends are of interest.

A different factor that is suspected to distort the international PISA rankings and trends is the varying effort that students invest into low-stake assessments. In the absence of individual feedback and direct consequences, students may not try their best and show less engagement in completing the test items correctly (Finn, 2015; Wise & Cotten, 2009). Results from former studies indicate an association between effort and test performance in low-stake assessments, which is likely to be moderated by country-level variables (Baumert & Demmrich, 2001; Gneezy et al., 2019; Pools & Monseur, 2021; Wise & DeMars, 2005).

To investigate, and possibly control for, a biasing effect of the level of effort on the test scores the PISA effort thermometer was constructed (Kunter et al., 2002). The effort thermometer is a self-report measure that is presented to students after they completed the assessment and is designed to quantify their perceived engagement during the test-taking process. Due to the subjective nature of self-reported effort measures their results must be interpreted with caution (Kong et al., 2007; Wise & Gao, 2017). Furthermore, given its placement at the end of the test-taking process, the effort thermometer may lend more weight towards the perceived engagement during the last part of the assessment (Pools & Monseur, 2021). Nevertheless it was shown that the scores from the effort thermometer do correlate with the students' proficiency level (Butler & Adams, 2007). In addition, both the average level of effort and the student-level correlation with the PISA scores appeared to differ across countries and subpopulations within countries (Butler & Adams, 2007; OECD, 2019a). This implies that country- and subpopulation-by-country-specific analyses are needed to evaluate the impact of effort on the differences between cycles in the national PISA trends.

The report is structured as follows: chapter 3 outlines the data analyzed in this study. In chapter 4 the functioning of trend items within the Netherlands is investigated across the three PISA cycles 2012, 2015 and 2018. In chapter 5 the national trends in reading, mathematics and science literacy are re-estimated for the Netherlands. Chapter 6 explores the relationship between self-perceived effort and student performance across countries and within the Netherlands. Finally, chapter 7 summarizes the results and provides advice for future research.

3. Data

The study is based on the Dutch PISA samples from 2012, 2015 and 2018. It encompasses the three domains reading literacy, mathematics literacy, and science literacy. All students in the 2012 sample took the paper-based form of the PISA assessment and all Dutch students in the 2015 and 2018 samples took the computer-based test.

Each PISA test consists of a mixture of newly developed items and some re-use of existing item material. These latter so called *trend items* are thus items that are administered in multiple PISA cycles and are necessary to perform a concurrent calibration (Hanson & Béguin, 2002) that is carried out to equate the results obtained on different test forms. Trend items therefore play a crucial role in order to achieve comparability across different cycles. Anchor items are here defined as trend items that were presented in the Netherlands across the three investigated cycles of the PISA study.

The number of trend and anchor items for which data from students in the Netherlands are available and included in this study is shown in Table 3.1.

	Number of Trend and Anchor Items			
PISA Domain	*2012*	*2015*	*2018*	*Anchor*
Reading	41	85	69	41
Mathematics	84	69	70	69
Science	53	83	115	39

Table 3.1: Number of trend and anchor items in the Netherlands

Table 3.1 shows that science and reading have about the same number of anchor items. For the mathematics domain, the number of anchor items is about twice as large (69). This is because the mathematics domain was redeveloped in 2012 (the first cycle included in this study), whereas the science and reading domains were redeveloped in 2015 and 2018 respectively.

Female or male students that were part of the VMBO GT, VMBO BB, VMBO KB, HAVO or VWO educational program are selected for the analyses. The number of students in the Netherlands for whom data are available, shown per item set and (sub)population is displayed in Table 3.2.

		NLD Samples for Trend and [Anchor] Items					
PISA Domain	**(Sub)population**	*2012*		*2015*		*2018*	
Reading	*Pooled*	2894	[2894]	2101	[1830]	4464	[4425]
	Female	1403	[1403]	1058	[921]	2203	[2185]
	Male	1491	[1491]	1043	[909]	2261	[2240]
	VMBO GT	805	[805]	605	[529]	1190	[1170]
	VMBO BB+KB	655	[655]	470	[412]	1185	[1172]
	HAVO	753	[753]	534	[469]	1101	[1096]
	VWO	681	[681]	492	[420]	988	[987]
Mathematics	*Pooled*	4213	[4211]	2107	[2107]	2669	[2669]
	Female	2053	[2052]	1104	[1104]	1292	[1292]
	Male	2160	[2159]	1003	[1003]	1377	[1377]
	VMBO GT	1169	[1168]	603	[603]	662	[662]
	VMBO BB+KB	946	[945]	486	[486]	895	[895]
	HAVO	1086	[1086]	575	[575]	583	[583]
	VWO	1012	[1012]	443	[443]	529	[529]
Science	*Pooled*	2902	[2899]	3013	[1731]	2710	[2710]
	Female	1416	[1415]	1542	[883]	1345	[1345]
	Male	1486	[1484]	1471	[848]	1365	[1365]
	VMBO GT	794	[793]	865	[496]	651	[651]
	VMBO BB+KB	662	[661]	682	[391]	923	[923]
	HAVO	755	[754]	796	[453]	603	[603]
	VWO	691	[691]	670	[391]	533	[533]

Table 3.2: Number of students in the Netherlands per item set and (sub)population

The *Pooled* population in Table 3.2 refers to the complete sample that includes the available data from all subpopulations. To achieve a sufficient sample size the VMBO BB and VMBO KB groups were merged into the *VMBO BB+KB* group. Furthermore, for the DIF analysis the VMBO GT, VMBO BB and VMBO KB groups were merged into the *VMBO* group, and the HAVO and VWO groups were merged into the *HAVO+VWO* group.

Table 3.2 displays that all subpopulations have at least 400 observations. The lowest number students is the VWO group in 2015 who have administered mathematics (443). All other groups have more observations. These numbers are sufficient for carrying out the subsequent analyses robustly. The highest number of students within each cycle reflects the main domain within each cycle (Mathematics in 2012; Science in 2015; Reading in 2018).

For the differential item functioning analysis (presented in Chapter 4) only students with responses to the anchor items are included. The trend analysis (Chapter 5) is moreover based on students for whom responses to the trend items are available.

4. Differential Functioning of Anchor Items

4.1 Method

Each PISA assessment contains a number of items from the previous cycles. These trend items serve as an anchor to establish a common scale for the ability estimates of students from different PISA cycles. Thereby the trends in reading, mathematics and science ability can be investigated. The focus of this study lies on the trends between 2012, 2015 and 2018. Trend items that were presented in all three cycles are denoted as anchor items and are investigated for differential item functioning (DIF) (Walker, 2011).

In this study an item is flagged as functioning differently if the probability of giving a particular response depends not only on the students' ability level but also on the PISA cycle of which they were part. For example, students in 2015 may be less likely to respond correctly to a certain item than students with the same underlying ability level in 2018. DIF in anchor items therefore threatens the comparability of ability estimates across cycles and can lead to false conclusions about the corresponding trends.

Two types of DIF are investigated. Uniform DIF describes the case where an item is consistently more difficult for students of one PISA cycle compared to another cycle. The gap in difficulty does therefore not depend on the underlying ability level and remains constant across the entire proficiency spectrum. In contrast, non-uniform DIF describes the case where the difference in difficulty between cycles varies depending on the students' ability level. For example, for students at the higher end of the proficiency spectrum an item may be more difficult in 2015 than in 2018, while at the lower end the difference is nonexistent or reversed.

The absence of uniform DIF is referred to as scalar invariance and implies that the difficulty parameters of the anchor items can be constrained to be equal across PISA cycles. Similarly, the absence of non-uniform DIF implies that the corresponding factor loadings, or discrimination parameters, do not vary between between cycles and is referred to as metric invariance.

A stepwise procedure for measurement invariance testing is applied to detect items that violate metric and/or scalar invariance (Putnick & Bornstein, 2016). In the first step, items that show non-uniform DIF between 2015 and 2018 are identified. In the second step, items with uniform DIF between 2015 and 2018 are identified while controlling for the DIF found in the first step. The procedure is repeated to compare item functioning between the paper-based (2012) and computer-based (2015+2018) groups while controlling for the DIF found within the computer-based group.

Inferences are made on a .05 significance level. A Bonferroni correction is applied to control for an inflated type-1 error rate due to the multiple comparisons made in the stepwise procedure (Armstrong, 2014). The DIF analyses are carried out using the R packages Lavaan and semTools (Jorgensen et al., 2021; R Core Team, 2021; Rosseel et al., 2021) and are based on the scaled chi-squared difference test by Satorra and Bentler.

Note that the sensitivity of the test statistic increases with sample size (Tong & Bentler, 2013). It is therefore expected to find the largest proportions of DIF items in the pooled groups. Evidence for subpopulation-specific DIF arises when items are flagged in a subpopulation but not in the pooled sample, or if the type of DIF (uniform/non-uniform) found differs from the pooled sample. Whether or not the DIF represents a systematic effect on item parameters (e.g., math items are consistently more difficult for male students in one cycle) or are the result of random variation (DIF effects cancel out) is explored in Chapter 5.

The results from the DIF analyses for the reading domain between 2015 and 2018 must be interpreted with caution. The Full Information Maximum Likelihood (FIML) estimation method does not condition on the Multistage Adaptive Test (MSAT) design deployed for the reading domain at the 2018 cycle, which can lead to biased item parameters for that group (Zwitser & Maris, 2015). As a consequence, reading items can be falsely flagged as functioning differently between 2015 and 2018. The false flags can lead to an inflated proportion of detected DIF items but do not affect the trend estimates in Chapter 5, as falsely flagged DIF items are by definition equally difficult (or discriminative) across groups.

In contrast, the fixed parameter linking approach applied for trend items in the 2018 PISA study can be a source of systematic DIF. In 2018 the trend items were not rescaled but fixed to the values obtained at the 2015 cycle (OECD, 2019b, Chapter 9). It has been shown that adaptive and sequential test designs can interact differently with student-level variables that are related to test performance, such as motivation, commitment or test anxiety (e.g., Asseburg & Frey, 2013; Kimura, 2017; Ling et al., 2017; Martin & Lazendic, 2018). It is therefore plausible that DIF detected in reading items between 2015 and 2018 cannot be fully attributed to country-specifc DIF but may also reflect the change in test design.

Finally, an anchor item that was detected as functioning differently between 2015 vs. 2018 could not be flagged again for the same type of DIF in the paper- vs. computer-based analysis. This follows from the sequential nature of the DIF analysis and is in line with the overarching goal to establish sets of anchor items that allow for a fair comparison of performance across the 2012, 2015 and 2018 PISA cycles.

4.2 Reading

4.2.1 Results

In Table 4.1 the number and relative percentages of anchor items for the reading domain with statistical significant differences (DIF) within the Netherlands between 2015 and 2018 are presented[1].

	Number of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Access and retrieve	5	0	5	2	1
Integrate and interpret	16	9	11	7	6
Reflect and evaluate	5	3	4	2	1
Open Response - Human Coded	13	7	11	4	3
Open Response - Computer Scored	1	0	0	0	0
Complex Multiple Choice	4	0	3	3	2
Simple Multiple Choice	8	5	6	4	3
Overall	26	12	20	11	8
	Percentage of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Access and retrieve	50	0	50	20	10
Integrate and interpret	70	39	48	30	26
Reflect and evaluate	62	38	50	25	12
Open Response - Human Coded	68	37	58	21	16
Open Response - Computer Scored	50	0	0	0	0
Complex Multiple Choice	50	0	38	38	25
Simple Multiple Choice	67	42	50	33	25
Overall	63	29	49	27	20

Table 4.1: Reading: DIF between 2015 and 2018 summarized

[1] The complete results can be found in the Appendix.

Within the next step, we have evaluated the consistency of the measurement between 2012 versus 2015-2018 combined. This is more or less the evaluation if the change in administration mode - from a paper based assessment mode in 2012 to a computer based assessments mode starting from 2015 - has led to differential item functioning. These results can be found in Table 4.2.

	Number of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Access and retrieve	8	3	8	8	5
Integrate and interpret	20	12	16	14	10
Reflect and evaluate	6	4	4	2	3
Open Response - Human Coded	17	11	14	12	7
Open Response - Computer Scored	2	1	2	2	1
Complex Multiple Choice	6	2	6	3	5
Simple Multiple Choice	9	5	6	7	5
Overall	34	19	28	24	18

	Percentage of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Access and retrieve	80	30	80	80	50
Integrate and interpret	87	52	70	61	43
Reflect and evaluate	75	50	50	25	38
Open Response - Human Coded	89	58	74	63	37
Open Response - Computer Scored	100	50	100	100	50
Complex Multiple Choice	75	25	75	38	62
Simple Multiple Choice	75	42	50	58	42
Overall	83	46	68	59	44

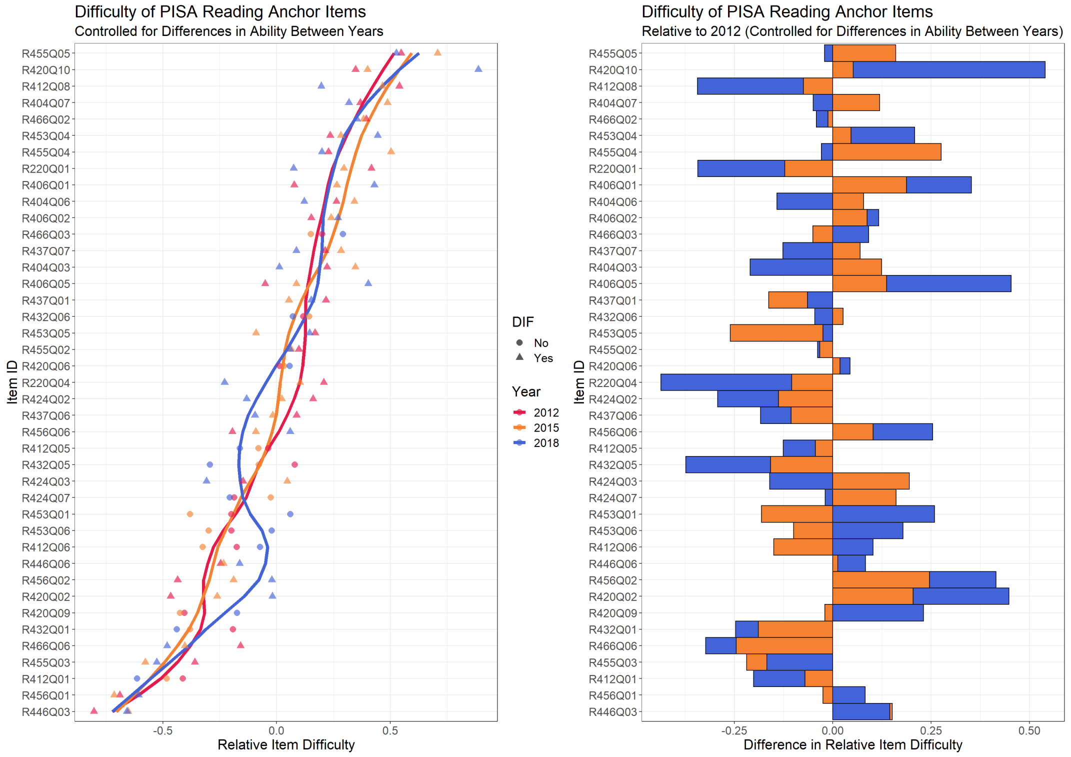
Table 4.2: Reading: DIF between PBA and CBA summarized

Across genders and educational programs, 83% of the reading items were flagged as functioning differently between the 2012, 2015 and 2018 PISA cycles. 63% of the anchor items were flagged with DIF in the 2015 vs. 2018 analysis while an additional 20% were flagged when comparing item functioning between the paper-based (2012) and the computer-based groups (2015 and 2018). The high rate of detected DIF items between 2015 and 2018 is a possible sign of falsely flagged DIF due to biased estimates of the 2018 reading item parameters.

Gender-specific analyses revealed a higher proportion of DIF items in the male group compared to the female group (Female: 46%; Male: 68%). The most noticeable differences were found for the response modes Open Response - Human Coded (Female: 58%; Male: 74%) and Complex Multiple Choice (Female: 25%; Male: 75%). Within the computer-based group, the DIF for male students could be fully attributed to shifts in item difficulty, while a mix of uniform and non-uniform DIF was detected for the female students.

The proportion of DIF items within the computer-based group was higher for VMBO than for HAVO+VWO students (VMBO: 27%; HAVO+VWO: 20%). The gap increased when including DIF between the paper- and computer-based groups (VMBO: 59%; HAVO+VWO: 44%). However, for Complex Multiple Choice items a higher proportion of DIF was detected in the HAVO+VWO group (VMBO: 38%; HAVO+VWO: 62%).

Within Figure 4.1a the relative difficulty of anchor items for the reading domain is shown. A positive/negative difficulty indicates that the item is harder/easier than the average difficulty within the year.

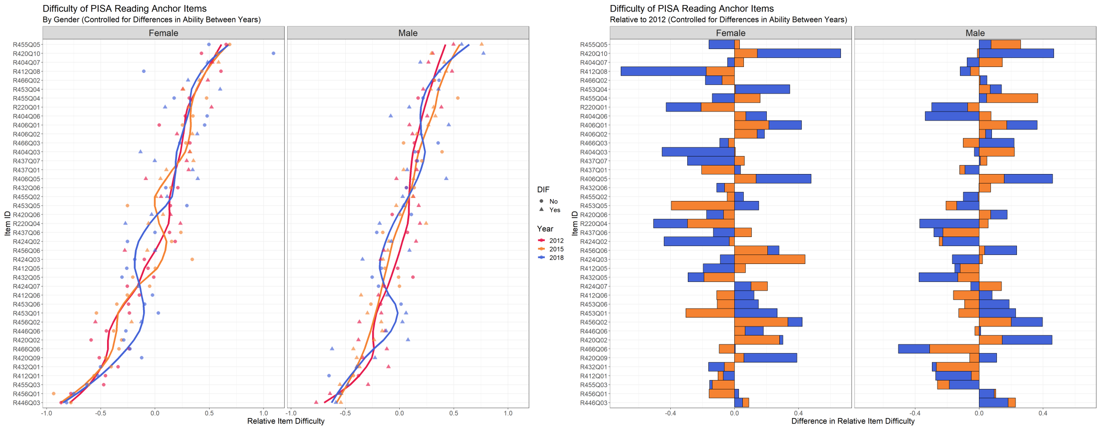


(a) Relative difficulty of anchor items (b) Change in relative difficulty compared to 2012

Figure 4.1: Reading: Relative difficulty of anchor items

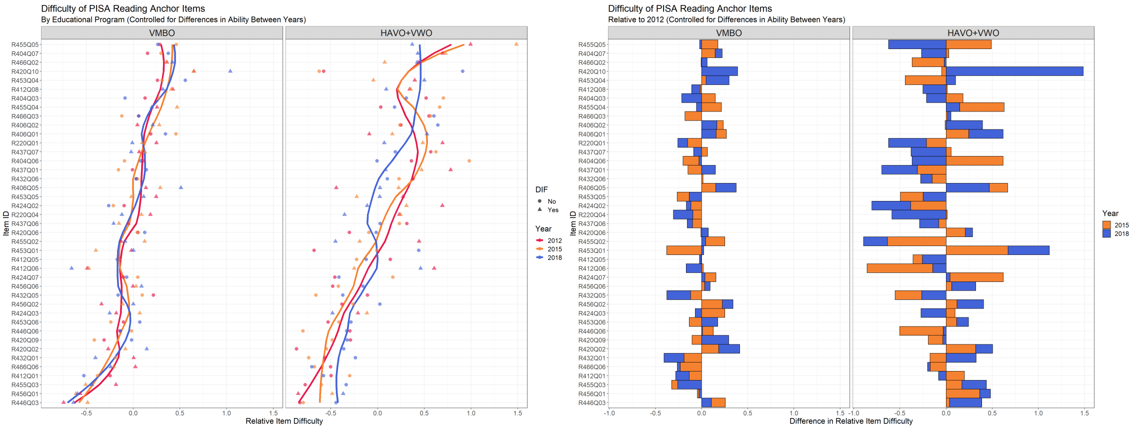
Most items that became relatively easier or harder in 2015 compared to the paper-based cycle showed the same pattern in 2018, as illustrated in Figure 4.1b.

The shift in relative difficulty was largely consistent across genders (Figure 4.2), but not educational programs (Figure 4.3).



(a) Relative difficulty of anchor items (b) Change in relative difficulty compared to 2012

Figure 4.2: Reading: Relative difficulty of anchor items by gender



(a) Relative difficulty of anchor items

(b) Change in relative difficulty compared to 2012

Figure 4.3: Reading: Relative difficulty of anchor items by educational track

In 2012 and 2015, the variation in relative difficulty was lower for the VMBO group compared to the HAVO+VWO group, i.e., for the VMBO students the difficulty level was more consistent across items (Figure 4.3a). Furthermore, the difference between easy and hard items was more pronounced for HAVO+VWO students. The discrepancies between the educational groups were reduced at the 2018 PISA cycle, which can indicate that the adaptive test design was effective in handling the spread in reading ability within the HAVO+VWO group. However, given the possible bias in the reading item parameters at the 2018 cycle, it cannot be concluded that the findings can be attributed to systematic differences in item functioning between test designs or cycles.

The sets of items that became easier or harder showed only sparse overlap for the VMBO and HAVO+VWO groups. However, within an educational group, the direction of the shift in difficulty was consistent between the 2015 and 2018 cycles (Figure 4.3b).

4.2.2 Summary

For 83% of the anchor items in the reading domain uniform and/or non-uniform DIF was detected. 63% of the anchor items were flagged with DIF in the 2015 vs. 2018 analysis while an additional 20% were flagged when comparing item functioning between the paper-based (2012) and the computer-based groups (2015 and 2018). It is suspected that biased item parameters at the 2018 cycle inflated the proportion of items flagged with DIF between 2015 and 2018.

Subpopulation-specific analyses showed a possible effect of the adaptive test design on the difficulty parameters of reading items in 2018. The effect was primarily observed in the HAVO+VWO group, which can be indicative of a program-test design or ability-test design interaction that threatens the comparability of item parameters across the PISA cycles. However, while a suspected inflation of DIF items between 2015 and 2018 does not affect the trend estimates, it does distort the results of the DIF analyses for the reading domain. The findings therefore do not pose comprehensive evidence in favour of an interaction between item functioning and changes in the test design and require further investigation.

4.3 Mathematics

4.3.1 Results

In Table 4.3 the summarized results of the DIF analysis for mathematics can be found [2].

	Number of Anchor Items with DIF				
Item Group	*All*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Space and Shape	3	2	1	0	1
Quantity	2	3	2	3	1
Uncertainty and Data	4	1	2	2	4
Change and Relationships	4	1	1	2	0
Open Response - Human Coded	2	3	1	1	1
Open Response - Computer Scored	8	2	3	4	2
Complex Multiple Choice	1	0	1	1	2
Simple Multiple Choice	2	2	1	1	1
Overall	13	7	6	7	6

	Percentage of Anchor Items with DIF				
Item Group	*All*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Space and Shape	18	12	6	0	6
Quantity	11	17	11	17	6
Uncertainty and Data	22	6	11	11	22
Change and Relationships	25	6	6	12	0
Open Response - Human Coded	11	17	6	6	6
Open Response - Computer Scored	36	9	14	18	9
Complex Multiple Choice	8	0	8	8	15
Simple Multiple Choice	12	12	6	6	6
Overall	19	10	9	10	9

Table 4.3: Mathematics: DIF between 2015 and 2018 summarized

Similar levels of DIF items were detected across the four subpopulations when investigating item functioning between the individual computer-based groups (Female: 10%; Male: 9%; VMBO: 10%; HAVO+VWO: 9%).

Between genders, the difference in the proportion of DIF items was consistent across the measured aspects of mathematics literacy. In contrast, for the Space and Shape and Quantity aspects more items were flagged with DIF in the HAVO+VWO group while for the Uncertainty and Data and Change and Relationships aspects the proportion of flagged DIF items was higher in the VMBO group (Table 4.3).

[2] The items M943Q02 (2012: 0.3%; 2015: 0%; 2018: 0.7% correct responses) and M992Q03 (2012: 0.6%; 2015: 0%; 2018: 0% correct responses) did not have correct responses for the VMBO students at one or more PISA cycles and were excluded from the VMBO-specific DIF analysis.

In the paper- vs. computer-based comparisob more items were flagged for the male and HAVO+VWO groups (Female: 20%; Male: 29%;VMBO: 25%; HAVO+VWO: 33%, see Table 4.4)

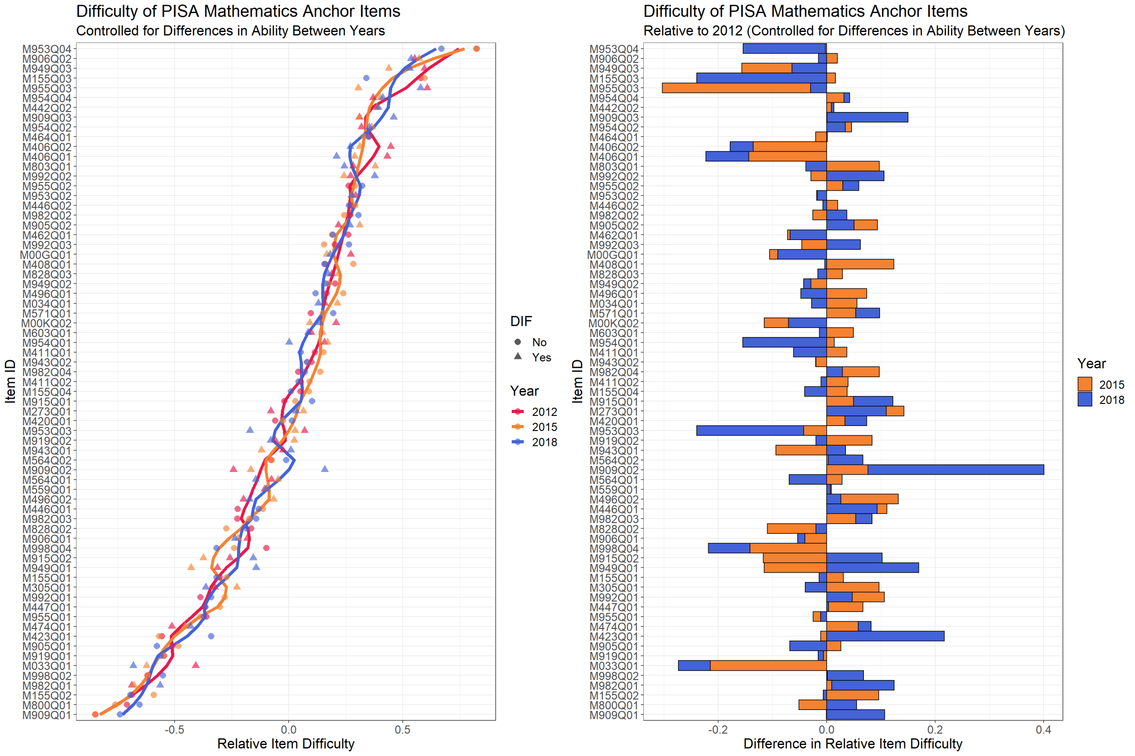
	Number of Anchor Items with DIF				
Item Group	*All*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Space and Shape	11	9	9	3	10
Quantity	12	7	8	8	9
Uncertainty and Data	6	2	4	8	7
Change and Relationships	6	3	5	5	3
Open Response - Human Coded	8	7	6	7	7
Open Response - Computer Scored	14	6	9	5	10
Complex Multiple Choice	4	3	4	6	5
Simple Multiple Choice	9	5	7	6	7
Overall	35	21	26	24	29

	Percentage of Anchor Items with DIF				
Item Group	*All*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Space and Shape	65	53	53	18	59
Quantity	67	39	44	44	50
Uncertainty and Data	33	11	22	44	39
Change and Relationships	38	19	31	31	19
Open Response - Human Coded	44	39	33	39	39
Open Response - Computer Scored	64	27	41	23	45
Complex Multiple Choice	31	23	31	46	38
Simple Multiple Choice	56	31	44	38	44
Overall	51	30	38	35	42

Table 4.4: Mathematics: DIF between PBA and CBA summarized

Across genders and educational programs, 51% of the mathematics items were flagged as functioning differently across 2012, 2015 and 2018 (Table 4.4. The proportion of DIF items was higher in the paper- vs. computer-based analysis (32%) compared to the 2015 vs. 2015 analysis (19%)

The relative difficulty of anchor items for the mathematics domain is found in Figure 4.4a. A positive/negative difficulty indicates that the item is harder/easier than the average difficulty within the year.

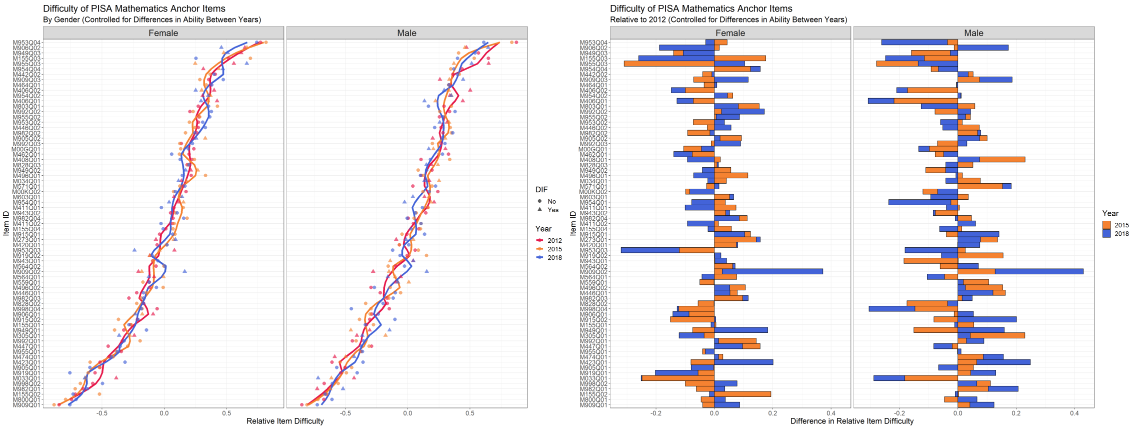


(a) Relative difficulty of anchor items

(b) Change in relative difficulty compared to 2012

Figure 4.4: Mathematics: Relative difficulty of anchor items

Most items that became relatively easier or harder in 2015 compared to the paper-based cycle showed the same pattern in 2018 (see Figure 4.4b) .



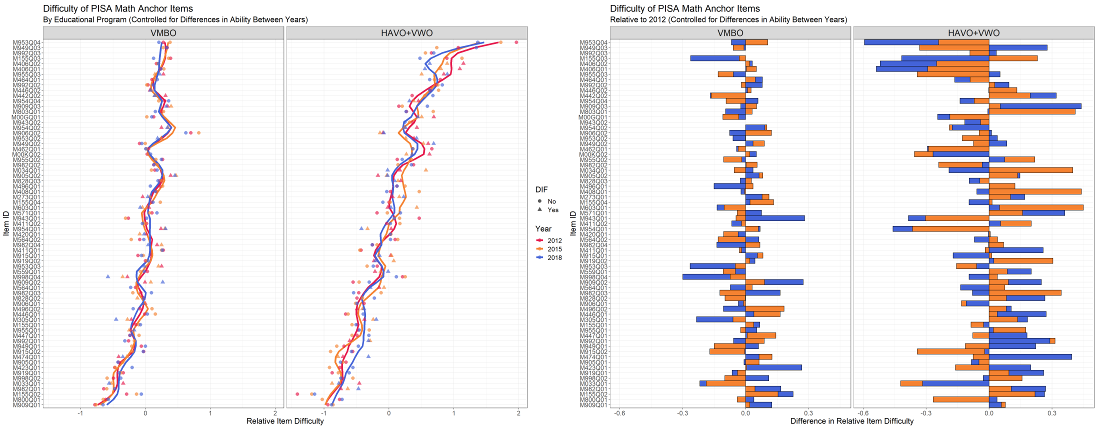
(a) Relative difficulty of anchor items

(b) Change in relative difficulty compared to 2012

Figure 4.5: Mathematics: Relative difficulty of anchor items by gender

As can be seen from Figure 4.5 the shift in relative difficulty was largely consistent across genders.

Less overlap was found between the sets of items that became easier (or harder) for the VMBO and HAVO+VWO groups (Figure 4.6a). However, within an educational group the item difficulties systematically shifted in the same direction for the 2015 and 2018 cycles (Figure 4.6b).



(a) Relative difficulty of anchor items

(b) Change in relative difficulty compared to 2012

Figure 4.6: Mathematics: Relative difficulty of anchor items by educational track

The difference in difficulty between the easiest and hardest items was less pronounced for the VMBO group compared to the HAVO+VWO group. As illustrated in Figure 4.6a, the finding was consistent across all PISA cycles.

4.3.2 Summary

Half of the anchor items in the mathematics domain (51%) were flagged as functioning differently across PISA cycles. The majority of DIF was detected when comparing responses between presentation modes. The results moreover indicate a possible program-specific mode effect. Within the computer-based group the level of DIF was consistent across genders and educational programs.

4.4 Science

4.4.1 Results

Across genders and educational programs, 62% of the science items were flagged as functioning differently between the 2012, 2015 and 2018 PISA cycles (Table . The proportion of DIF items was equally distributed between the paper- vs. computer-based analysis (31%) and the 2015 vs. 2018 analysis (31%).

	Number of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Explain phenomena scientifically	6	1	3	4	3
Evaluate and design scientific enquiry	5	4	3	2	2
Interpret data and evidence scientifically	1	2	1	0	1
Open Response - Human Coded	5	3	3	2	0
Open Response - Computer Scored	0	0	0	0	0
Complex Multiple Choice	4	3	2	3	5
Simple Multiple Choice	3	1	2	1	1
Overall	12	7	7	6	6

	Percentage of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Explain phenomena scientifically	46	8	23	31	23
Evaluate and design scientific enquiry	42	33	25	17	17
Interpret data and evidence scientifically	7	14	7	0	7
Open Response - Human Coded	45	27	27	18	0
Open Response - Computer Scored	0	0	0	0	0
Complex Multiple Choice	27	20	13	20	33
Simple Multiple Choice	23	8	15	8	8
Overall	31	18	18	15	15

Table 4.5: Science: DIF between 2015 and 2018 summarized

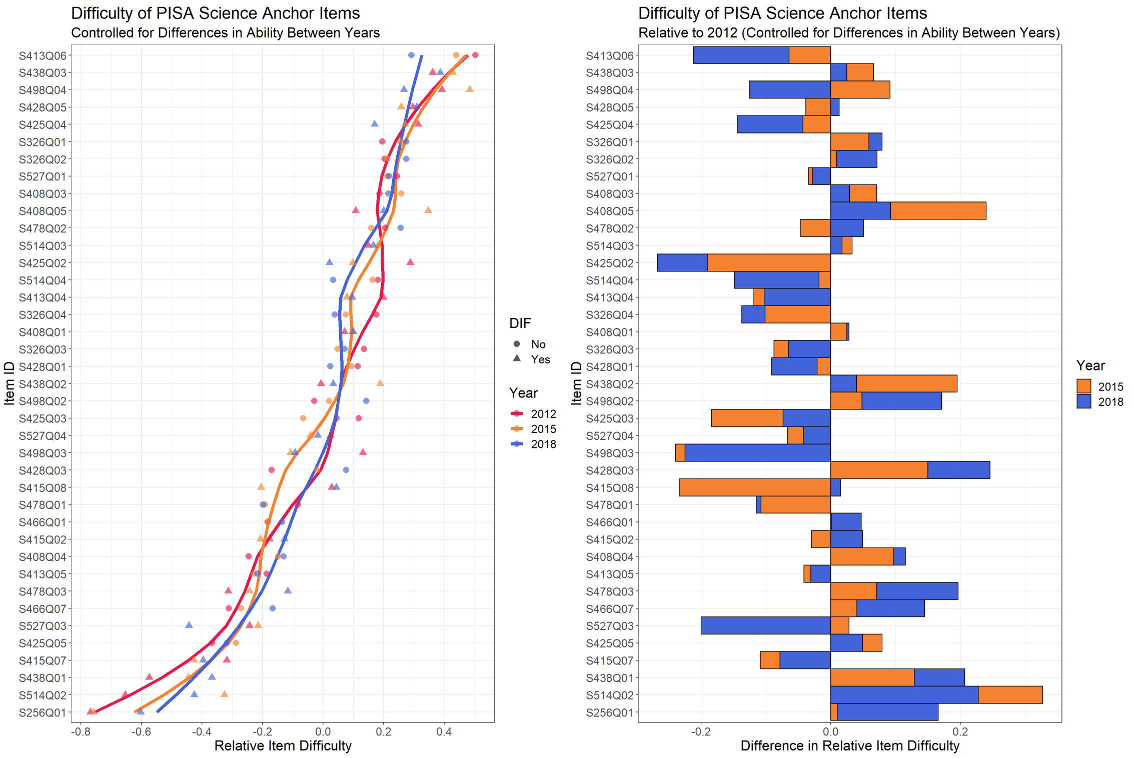
Within the computer-based group, the proportion of DIF items did not differ between genders or education programs (Female: 18%; Male: 18%; VMBO: 15%; HAVO+VWO: 15%). Except for one item flagged with non-uniform DIF in the female group, the DIF between the computer-based cycles could be fully attributed to items that showed an uniform shift in difficulty.

	Number of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Explain phenomena scientifically	8	4	5	7	6
Evaluate and design scientific enquiry	8	7	6	7	2
Interpret data and evidence scientifically	8	5	4	4	5
Open Response - Human Coded	7	5	5	4	2
Open Response - Computer Scored	0	0	0	0	0
Complex Multiple Choice	8	6	5	7	7
Simple Multiple Choice	9	5	5	7	4
Overall	24	16	15	18	13
	Percentage of Anchor Items with DIF				
Item Group	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
Explain phenomena scientifically	62	31	38	54	46
Evaluate and design scientific enquiry	67	58	50	58	17
Interpret data and evidence scientifically	57	36	29	29	36
Open Response - Human Coded	64	45	45	36	18
Open Response - Computer Scored	0	0	0	0	0
Complex Multiple Choice	53	40	33	47	47
Simple Multiple Choice	69	38	38	54	31
Overall	62	41	38	46	33

Table 4.6: Science: DIF between PBA and CBA summarized

In the paper- vs. computer-based comparison a higher number of DIF items was detected for the female and VMBO groups (Female: 23%; Male: 20%; VMBO: 31%; HAVO+VWO: 18%). For the educational groups, the discrepancy in the proportion of DIF items between presentation modes was rooted in the Evaluate and design aspect of scientific literacy (VMBO: 41%; HAVO+VWO: 0%). Moreover, both the sets of items flagged with DIF and the type of DIF found between the paper- and computer-based modes differed across the educational groups (Table A.6).

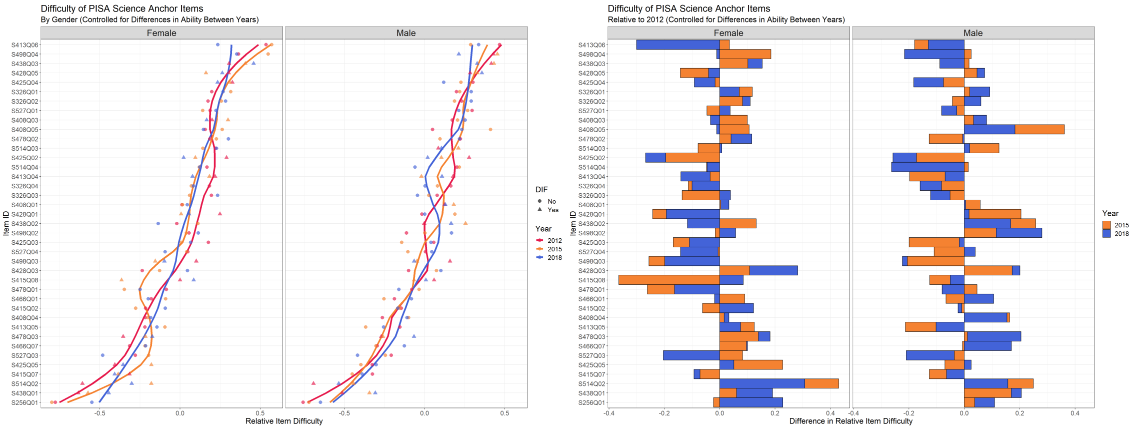
Items that saw a positive or negative shift in relative difficulty between 2012 and 2015 systematically showed a shift of same direction between 2012 and 2018 as displayed in Figure 4.7.



(a) Relative difficulty of anchor items (b) Change in relative difficulty compared to 2012

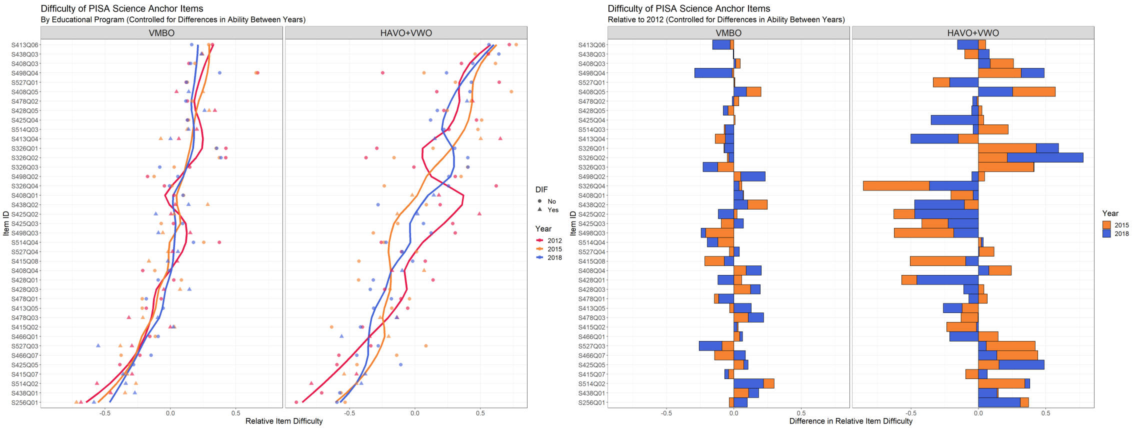
Figure 4.7: Science: Relative difficulty of anchor items

The direction of the shift varied between genders and educational groups, however the variation did not follow a systematic pattern and was spread evenly across easy and difficult items (Figures 4.8 and 4.9).



(a) Relative difficulty of anchor items (b) Change in relative difficulty compared to 2012

Figure 4.8: Science: Relative difficulty of anchor items by gender



(a) Relative difficulty of anchor items

(b) Change in relative difficulty compared to 2012

Figure 4.9: Science: Relative difficulty of anchor items by educational track

For HAVO+VWO students the item difficulty parameters were spread more widely and showed greater extreme values than for VMBO students. The pattern was visible across all PISA cycles but was most pronounced in 2012 (Figure 4.9a).

4.4.2 Summary

62% of the anchor items in the science domain were flagged as functioning differently across PISA cycles. The results indicate that the effect of presentation mode varied between students from different educational programs, but not between genders. The evidence for a program-mode interaction effect was most prominent for items that measure the Evaluate and design aspect of scientific literacy. Within the computer-based group no evidence of systematic DIF was observed.

5. Trend Estimates

5.1 Method

The trend in ability of students within the Netherlands across the 2012, 2015 and 2018 PISA cycles is estimated for the reading, mathematics and science domains. Three variants of the trend estimates are computed for each domain:

1. The first variant is based on the set of anchor items used in the PISA study and therefore does not control for DIF specific to the Netherlands.
2. The second variant controls for country-specific DIF of anchor items in the Netherlands.
3. In the case of analysing subpopulations (e.g., male and female students), a third variant is estimated that controls for country- and subpopulation-specific DIF of anchor items in the Netherlands.

For each variant a different set of anchor items establishes the common scale across years. The sets of anchor items chosen for the second and third variant are derived from the results of the DIF analysis shown in Tables A.1 and A.2 (reading), Tables A.3 and A.4 (mathematics), and Tables A.5 and A.6 (science). The anchor items for the second and third variant are therefore subsets of the anchor items used in the original PISA study. Anchor items that are flagged as functioning differently are treated as year-specific trend items and are estimated freely for each cycle.

The R package Dexter (Maris et al., 2021a; R Core Team, 2021) is used to obtain item parameters through the extended nominal response model (ENORM) and to subsequently draw 10 plausible values for each student conditional on their scored responses and the item parameters (Marsman et al., 2016). To account for the Multistage Adaptive Test (MSAT) design at the 2018 cycle, the item parameters for the reading domain are estimated with the Marginal Maximum Likelihood (MML) extension of Dexter (Maris et al., 2021b), which is an adequate choice under the assumption that the students' abilities follow a normal distribution (Steinfeld & Robitzsch, 2021).

The variation in plausible values represents the uncertainty in measuring the latent construct. In addition, replicate survey weights are applied when obtaining the trend estimates from the plausible values to account for the varying selection probability under the PISA sampling design (Lumley, 2021). The trend estimates for the first variant are scaled to match the three-year average and standard deviation of the plausible values from the original PISA study. The scaling does not affect the direction or magnitude of the differences between trend variants but allows visualizing the results on the common PISA score scale.

Within a domain, the ability estimates are comparable between years and subgroups. Moreover, the ability estimates for different variants share a common scale. Thereby the effect of controlling for country-specific DIF in anchor items can be examined by comparing the new trend line to the trend estimated under the original set of PISA anchor items. Divergent trends between variants suggest a systematic country-specific difference in item functioning for the Netherlands. For example, divergence in variants between 2012 and 2015 can indicate that the mode effect within the Netherlands of switching from paper-based to computer-based test forms differs from the global mode effect across all participating PISA countries.

In the following figures, the first, second and third trend variants are coded as red, blue and green lines, respectively. A dashed horizontal line indicates the three-year average within the Netherlands given the PISA anchor items for the analysed (sub)population. Vertical bars represent the 95% confidence intervals and express the uncertainty in the trend estimates.

5.2 Reading

5.2.1 Results

In Figure 5.1 globally and locally calibrated trends in reading ability within the Netherlands across the 2012, 2015 and 2018 PISA cycles are found.

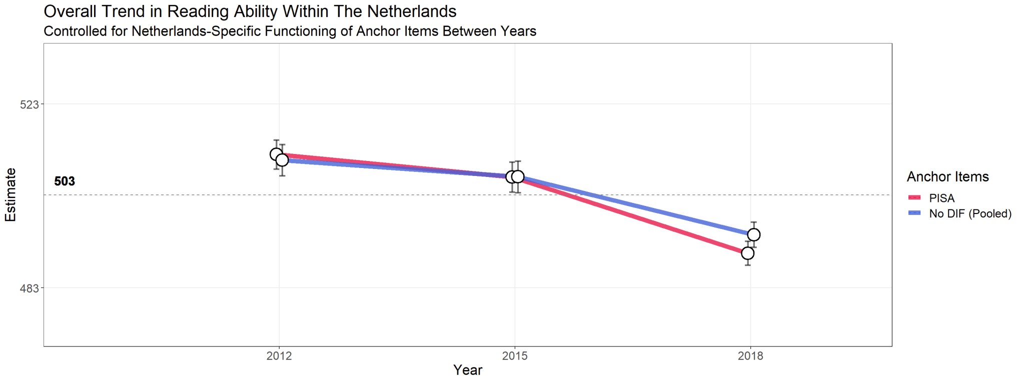


Figure 5.1: Trends in reading ability

Controlling for country-specific DIF in the Netherlands lead to a positive shift in the estimated ability level at the 2018 cycle, but did not affect the estimates at the 2012 and 2015 cycles (Figure 5.1).

Figures 5.2 and 5.3 show the trend results for males and females and different educational tracks using different calibrations designs within the Netherlands

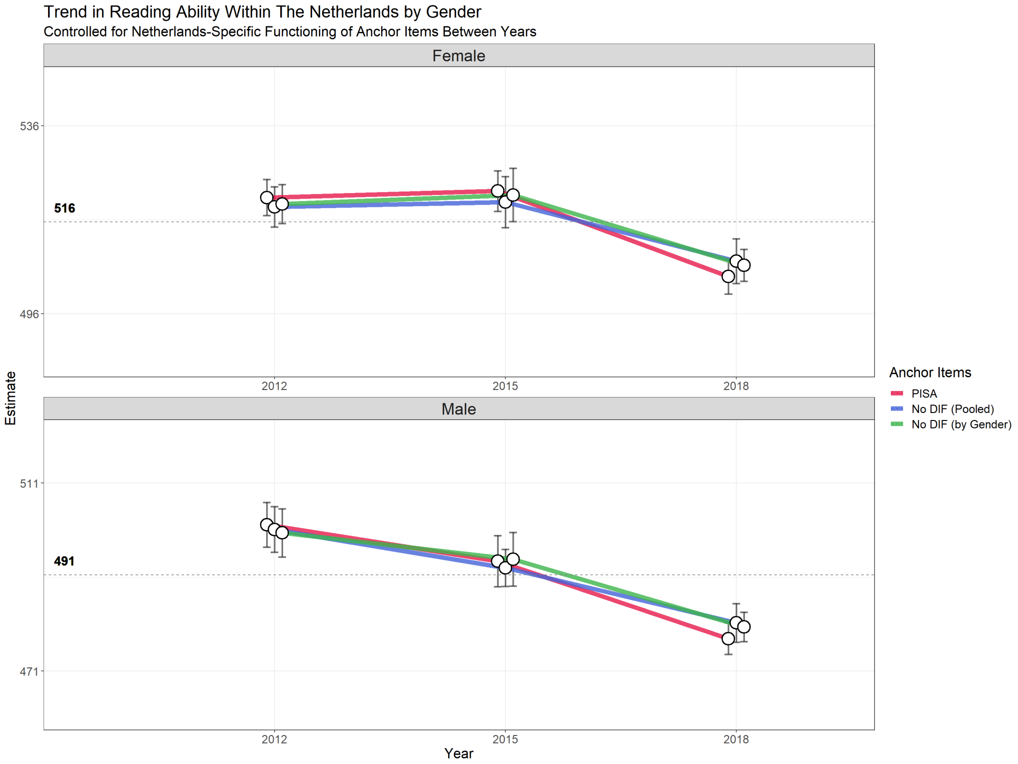


Figure 5.2: Trends in reading ability by gender

The pattern was equally pronounced for female and male students (Figure 5.2). In the program-specific analyses the differences between trend variants did not exceed the margin of error (Figure 5.3).

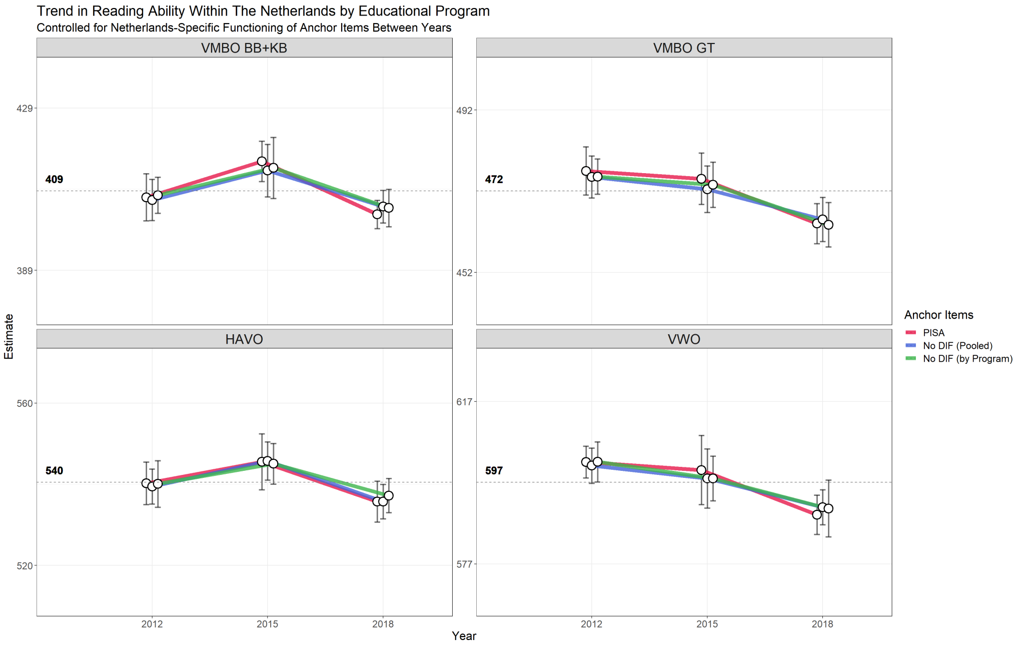


Figure 5.3: Trends in reading ability by educational track

It is important to note that the observed upward correction in reading performance at the 2018 cycle cannot be clearly attributed to country-specific DIF within the Netherlands. The fixed parameter linking approach applied for trend items in the 2018 PISA study comprises alternative explanations for a systematic difference between the original and the adjusted ability estimates (see Chapter 4 for an elaboration on the issue).

5.2.2 Summary

The estimated reading literacy within the Netherlands saw an upward correction at the 2018 PISA cycle when anchor items were controlled for country-specific DIF. The estimates at the 2015 and 2018 cycles were not affected by the local item calibration. It is unclear whether the shift in reading performance at the 2018 cycle can be attributed to Netherlands-specific DIF in anchor items, to an effect of the MSAT design on student variables (e.g., the level of motivation or commitment during the test-taking process), or to a mixture of both.

5.3 Mathematics

5.3.1 Results

The trend estimates were strongly consistent between the PISA variant and the variants that controlled for DIF specific to the Netherlands (Figure 5.4).

Controlling for country-specific DIF caused a slight upward correction in the mathematics ability level of VWO students at the 2015 cycle. The correction remained equally present when controlling for program-specific DIF. However, in both cases the divergence from the PISA variant was statistically small and could not clearly be distinguished from random measurement or sampling error (Figure 5.6).

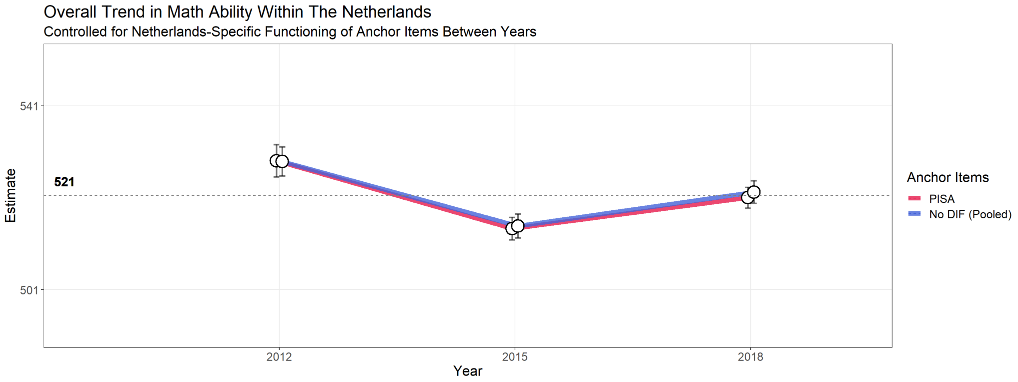


Figure 5.4: Trends in mathematics ability

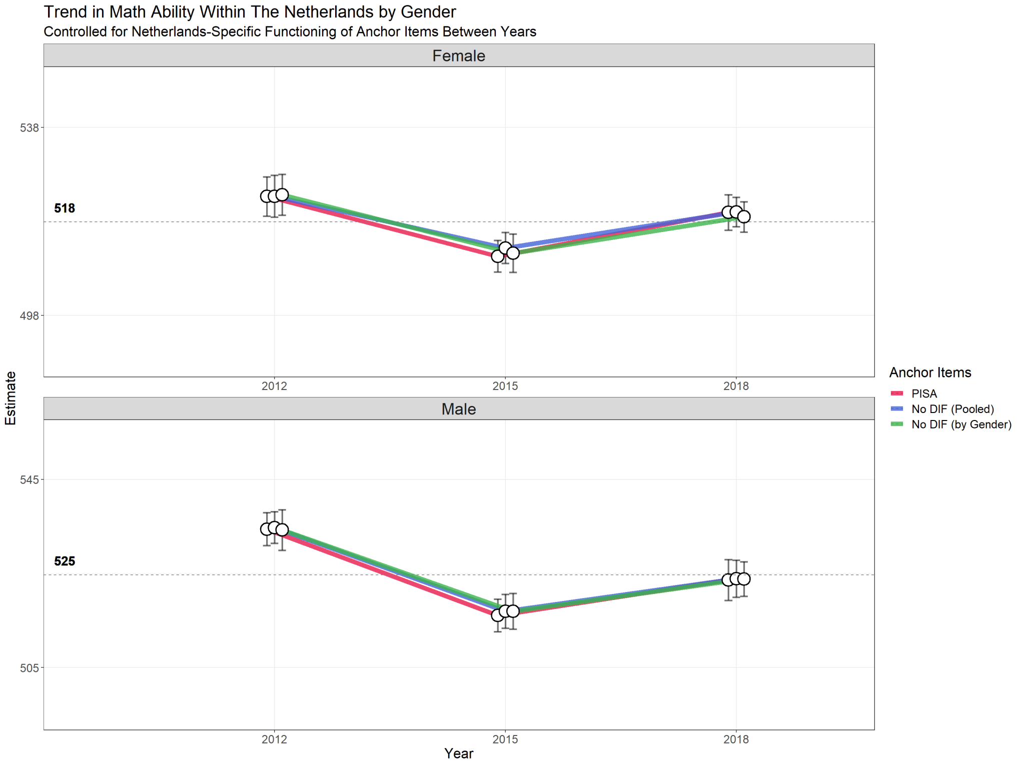


Figure 5.5: Trends in mathematics ability by gender

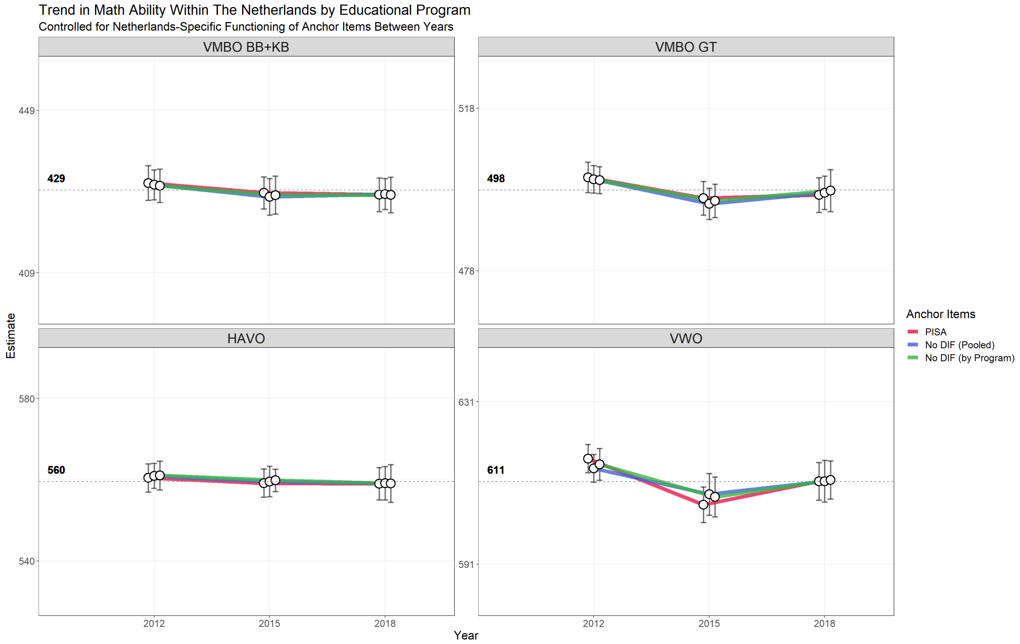


Figure 5.6: Trends in mathematics abilit by educational track

5.3.2 Summary

The results indicate that for the mathematics domain, a global calibration of anchor items is sufficient to capture possible item-mode and item-cycle interactions within the Netherlands. It was therefore not evident that the DIF detected in the analysis of data from the Netherlands reflects a systematic country-specific effect on the functioning of items. Instead, the DIF could be attributed to random variation in the data collection process.

5.4 Science

5.4.1 Results

The estimated ability level at the 2015 PISA cycle saw a strong downward correction when controlling for country-specific DIF in the Netherlands. Estimates at the 2012 and 2018 cycles remained stable and within the margin of error from the PISA trend (Figure 5.7).

The downward correction at the 2015 cycle was equally pronounced for female and male students and did not change when controlling for gender-specific DIF (Figure 5.8). However, when controlling for program-specific DIF only the VMBO groups were affected while for the HAVO and VWO groups no divergence from the (program-specific) PISA trends was observed (Figure 5.9).

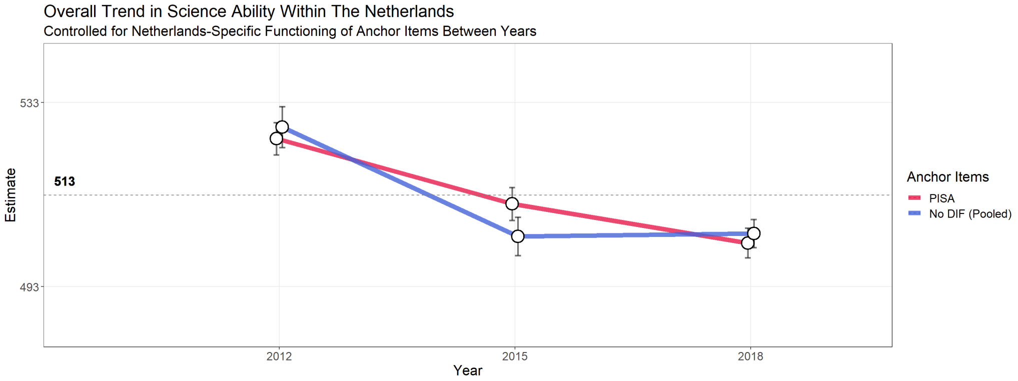


Figure 5.7: Trends in science ability

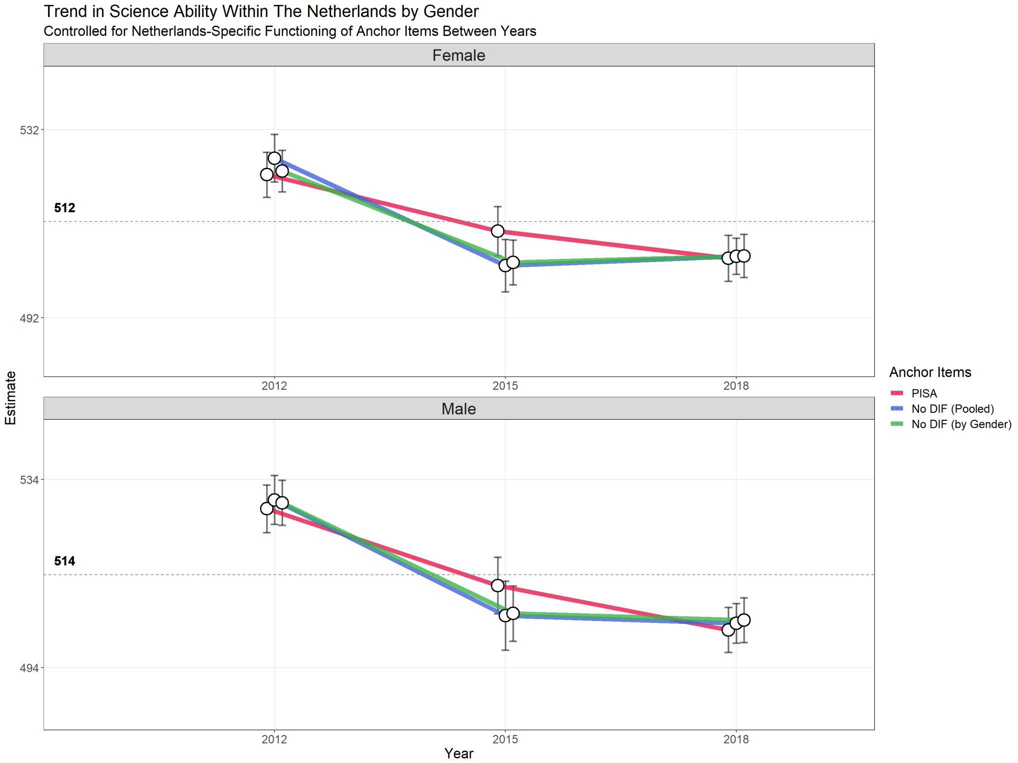


Figure 5.8: Trends in science ability by gender

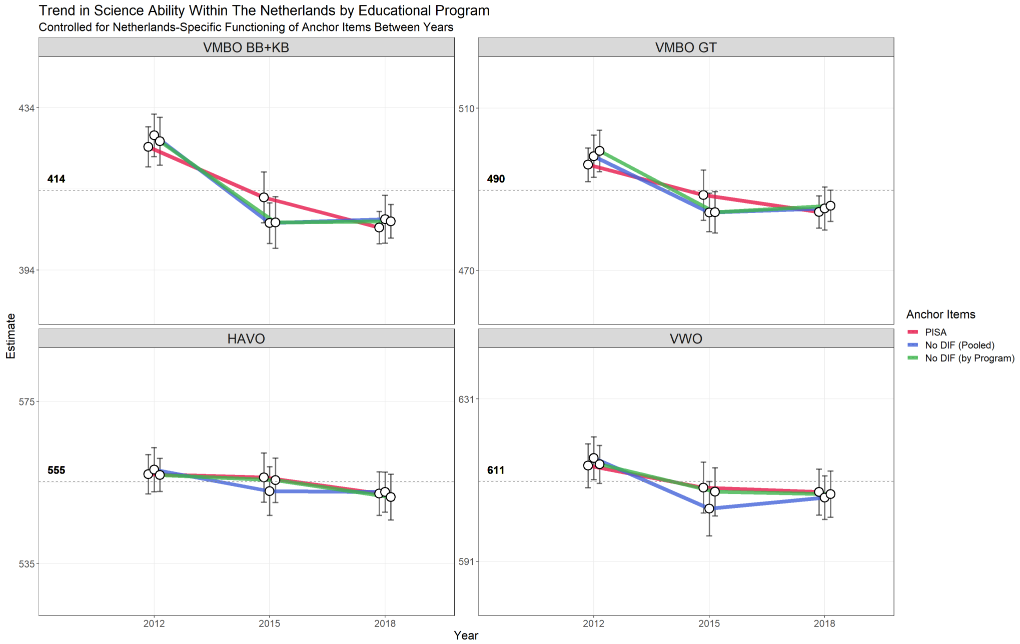


Figure 5.9: Trends in science ability by educational track

5.4.2 Summary

It is evident that for the 2015 PISA cycle, the global calibration of anchor items led to an overestimation of the scientific literacy within the Netherlands. The adjusted trends show a strong decline in ability between 2012 and 2015, and an equal level of performance across the computer-based cycles.

In program-specific analyses the drop in science performance at the 2015 cycle was observed for the VMBO groups, but vanished when controlling for local DIF in the HAVO and VWO groups. This implies that the country-specific DIF of anchor items was limited to the VMBO groups and that the global item parameters were adequate to measure the performance in the HAVO and VWO groups.

The results are moreover indicative of a mode effect within the Netherlands that explains the drop in science performance between the paper- and computer-based cycles. The suspected mode effect did not systematically differ between female and male students but showed an interaction with the followed educational program.

6. The Effect of Motivation

6.1 Method

The country-level averages of the effort scores from the 2018 PISA effort thermometer (Kunter et al., 2002) are compared to the shift in domain-specific performance of countries between the 2015 and 2018 cycles (OECD, 2019a). The effort scores represent student-level responses to the question *Compared to the situation you have just imagined, how much effort did you put into doing this test?* and were scored on a scale of 1 to 10, where higher numbers indicate a greater perceived engagement during the test-taking process. For the Netherlands-specific analyses, the correlation between the student-level effort scores and the students' domain-specific plausible values (as computed in the 2018 PISA study) is investigated. The analyses are carried out for a pooled sample across all subpopulations within the Netherlands, and furthermore split by gender and educational program.

6.2 International

6.2.1 Results

For the reading domain, a small positive correlation between the average student effort in 2018 and the change in country performance between 2015 and 2018 was found ($R = 0.16$). Given the reported level of invested effort, the decline in reading performance between 2015 and 2018 was stronger than expected for the Netherlands (Figure 6.1).

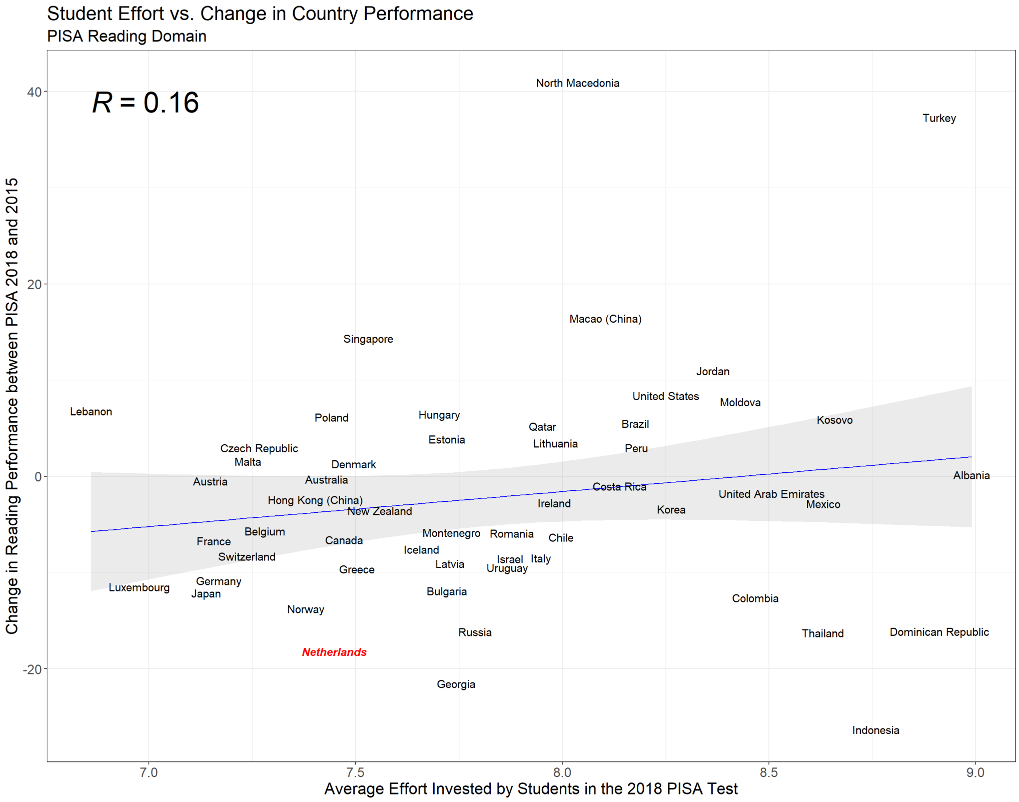


Figure 6.1: Correlation between effort and change in PISA reading scores

A moderate country-level correlation between student effort in 2018 and the change in mathematics performance between 2015 and 2018 was found ($R = 0.29$). In a comparison between countries, students in the Netherlands on average performed slightly better than expected at the 2018 cycle given their reported level of effort. (Figure 6.2).

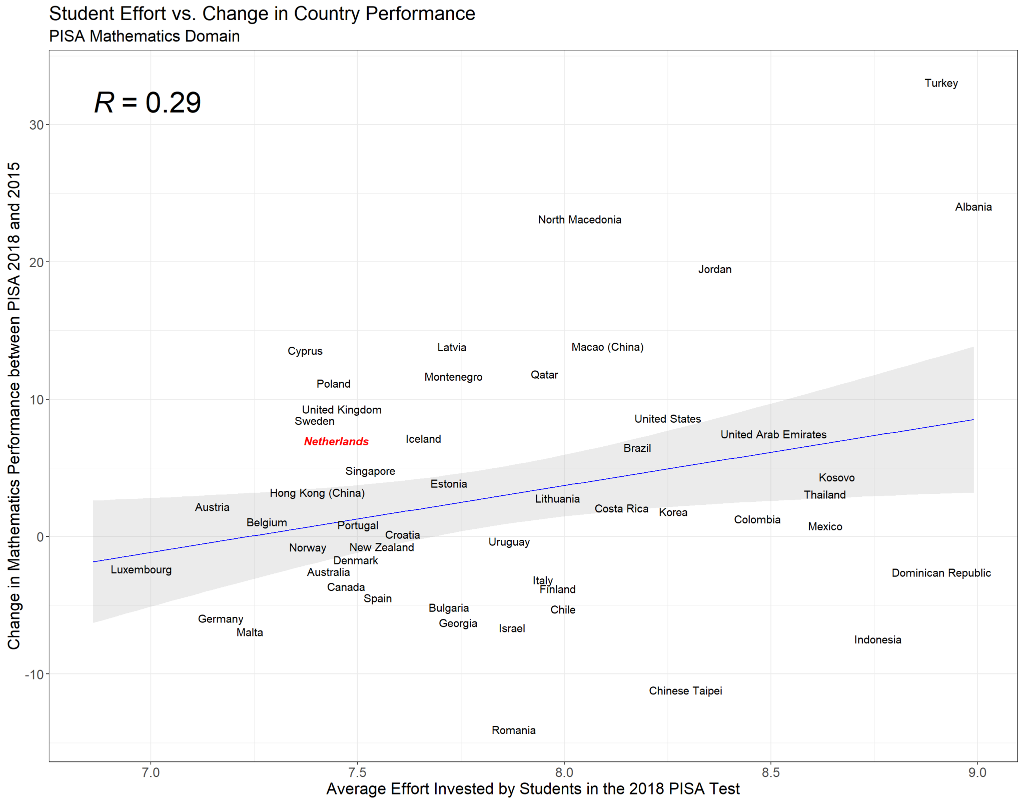


Figure 6.2: Correlation between effort and change in PISA mathematics scores

The country-level correlation between the change in science performance from 2015 to 2018 and the average reported effort at the 2018 cycle was moderately positive ($R = 0.28$). As illustrated in Figure 6.3, the performance of the Netherlands at the 2018 cycle was in line with the performance of other countries that reported a similar level of student effort.

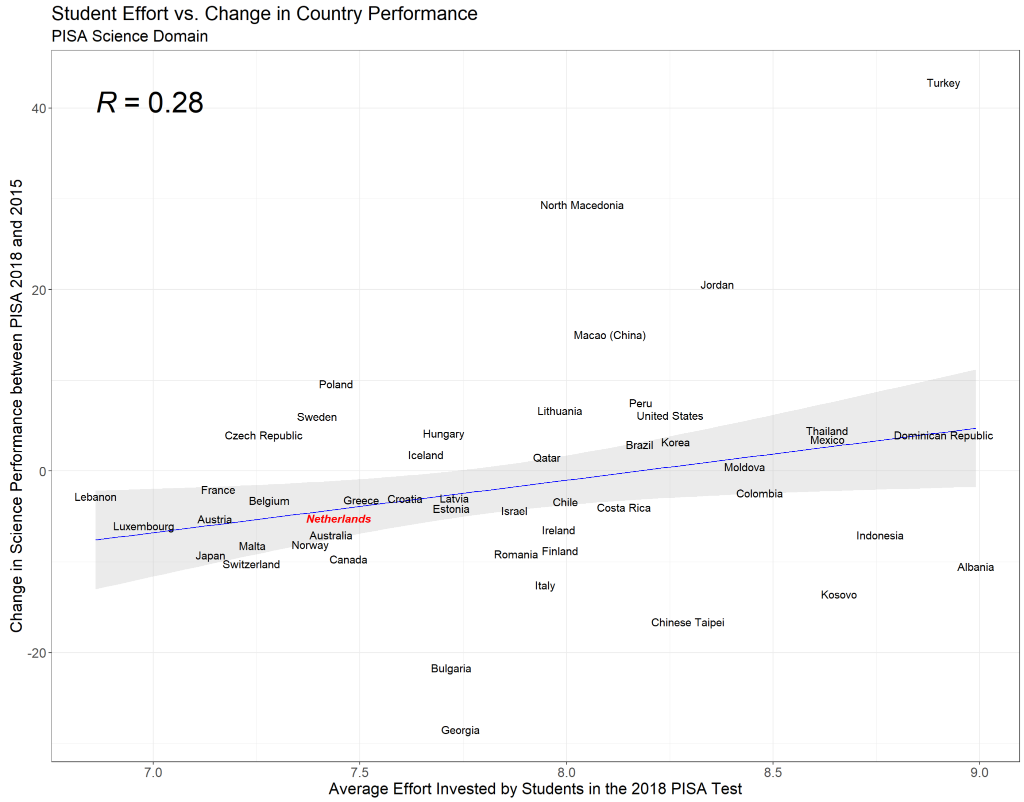


Figure 6.3: Correlation between effort and change in PISA science scores

The correlation between the average self-reported effort invested by students in the 2018 PISA test and the change in PISA reading scores between the 2018 and 2015 PISA cycles can be found in Figure 6.1. A positive change in scores indicates that a country performed better in 2018 compared to 2015 at the reading domain. The corresponding figures for mathematics and science can be found in Figures 6.2 and 6.3 respectively.

6.2.2 Summary

On country-level, student effort correlated positively with growth in performance between the 2015 and 2018 PISA cycles. The correlation was weak for the reading domain, and of moderate size for the mathematics and science domains. In the international comparison of reading performance at the 2018 cycle, students in the Netherlands on average performed worse than expected given their reported level of invested effort. For the mathematics and science domains, the performance of the Netherlands at the 2018 cycle did not noticeably differ from the performance of other countries that reported a similar level of student effort.

6.3 The Netherlands

6.3.1 Results

Figures 6.4, 6.5 and 6.6 show the correlations between the self-reported effort invested by students within the Netherlands and their estimated ability at the 2018 PISA cycle for respectively reading, mathematics and science. Darker color shades indicate that a higher number of students reported an effort level.

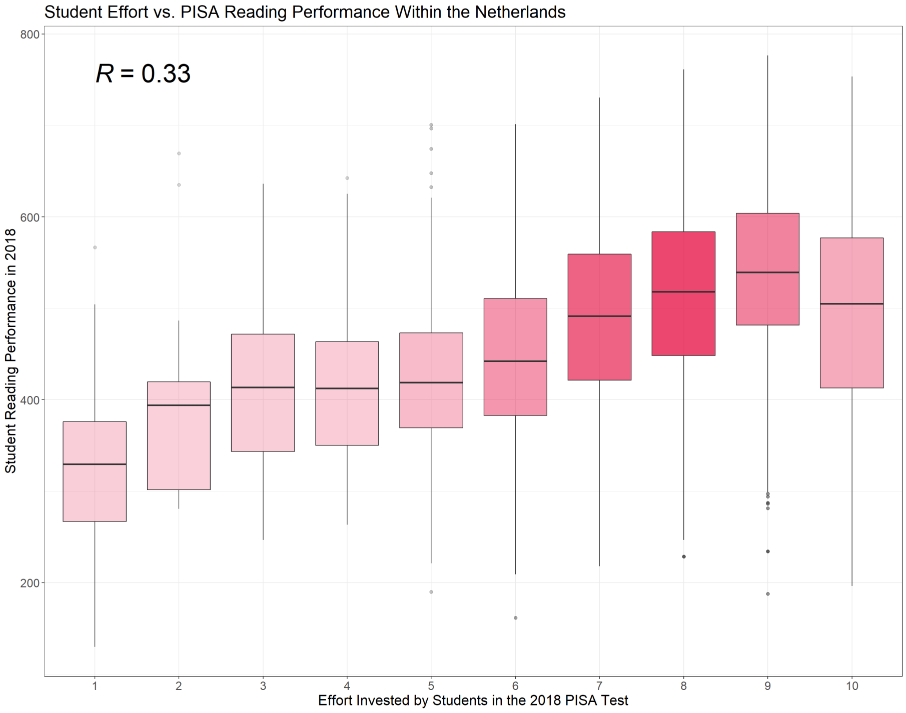


Figure 6.4: Correlation between effort and reading ability

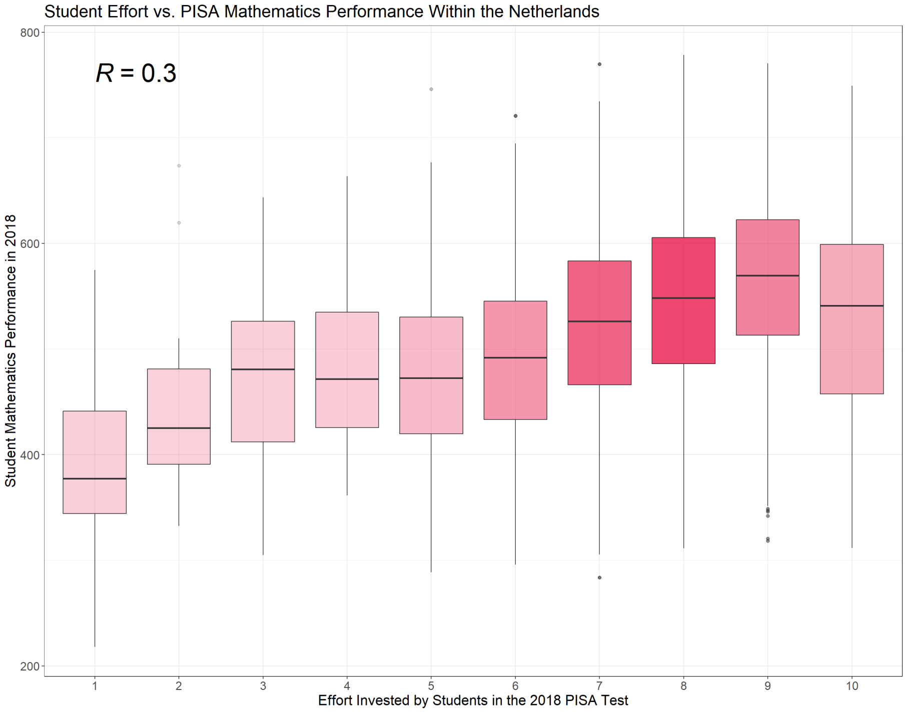


Figure 6.5: Correlation between effort and mathematics ability

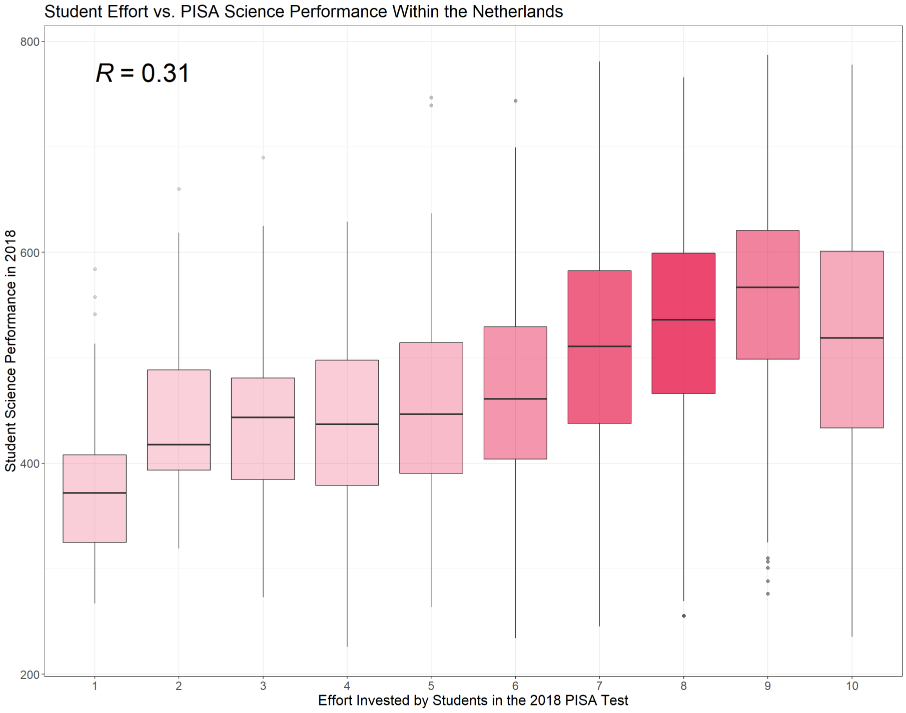


Figure 6.6: Correlation between effort and science ability

Across all domains, a country-specific analysis showed a moderate correlation between the reported effort of students in the Netherlands and their estimated level of proficiency (Reading: $R = 0.33$; Mathematics: $R = 0.30$; Science: $R = 0.31$). The majority of the students in the Netherlands reported an effort level between 6 and 9 on the PISA effort thermometer (Figures 6.4, 6.5 and 6.6).

The correlation between the self-reported effort invested by students within the Netherlands and their estimated ability at the 2018 PISA cycle, split by gender are presented in Figures 6.7, 6.8 and 6.9). Again, darker color shades indicate that a higher number of students reported an effort level.

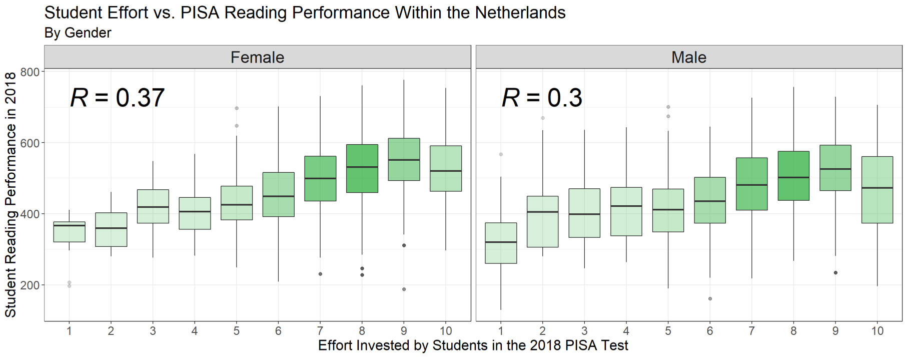


Figure 6.7: Correlation between effort and reading ability by gender

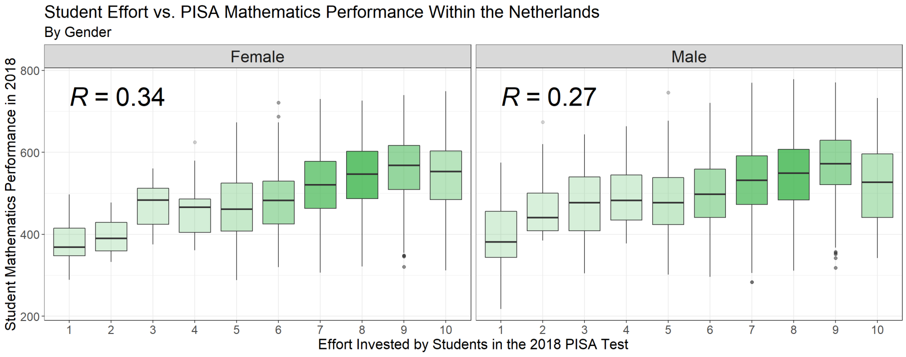


Figure 6.8: Correlation between effort and mathematics ability by gender

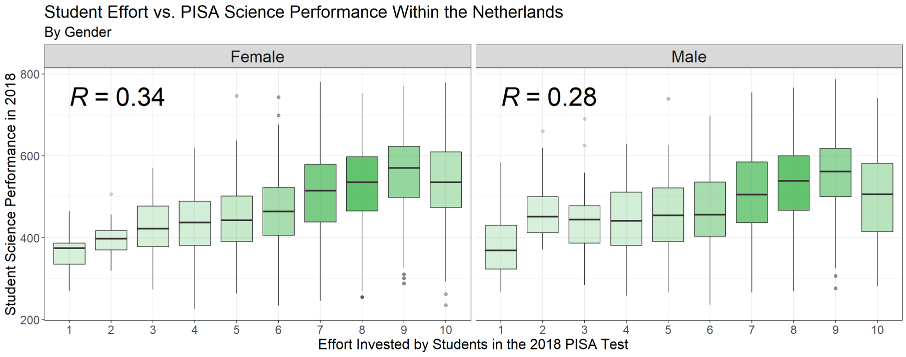


Figure 6.9: Correlation between effort and science ability by gender

The measured correlation between effort level and test performance was consistently stronger for female than for male students (Reading: $R_f = 0.37$, $R_m = 0.30$; Mathematics:

$R_f = 0.34$, $R_m = 0.27$; Science: $R_f = 0.34$, $R_m = 0.28$). The distribution of effort scores was coherent across genders

We also evaluated the relationship between effort and ability, split by the different educational programs found in the Netherlands. These results are found in Figures 6.10, 6.11 and 6.12).

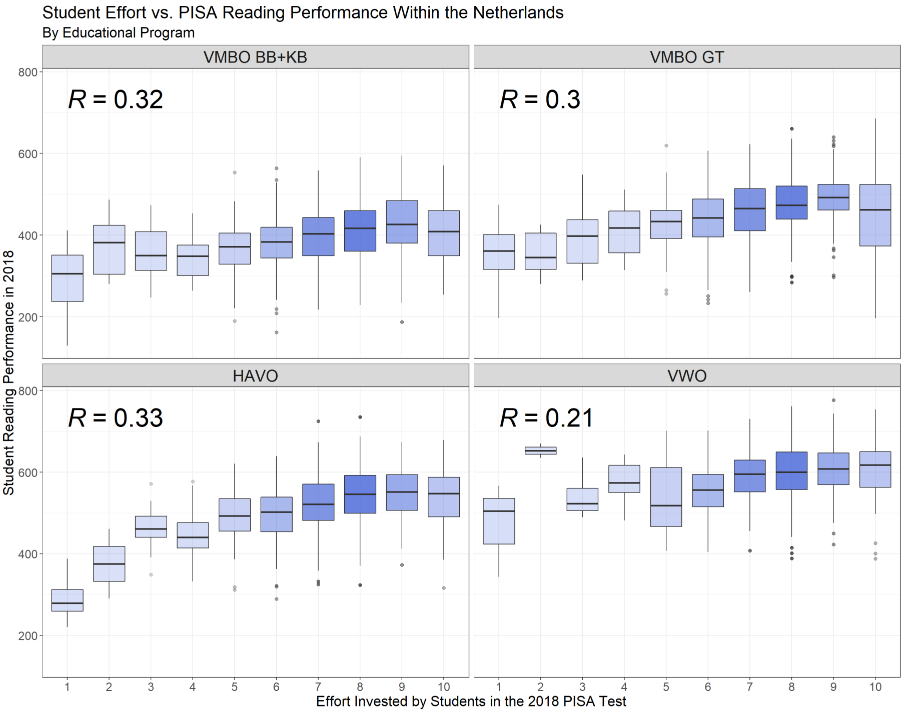


Figure 6.10: Correlation between effort and reading ability by educational track

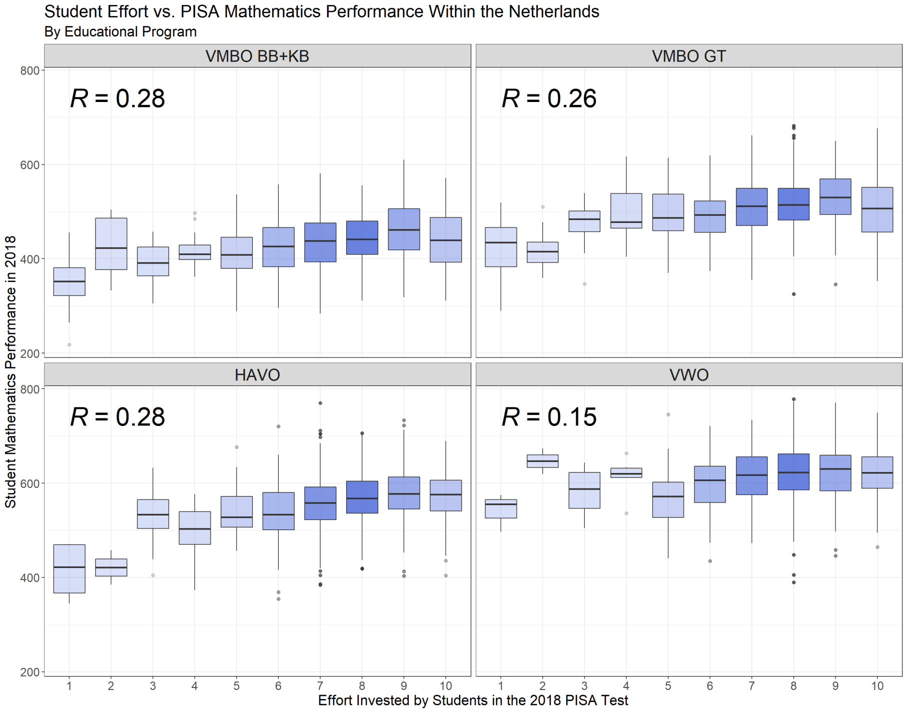


Figure 6.11: Correlation between effort and mathematics ability by educational track

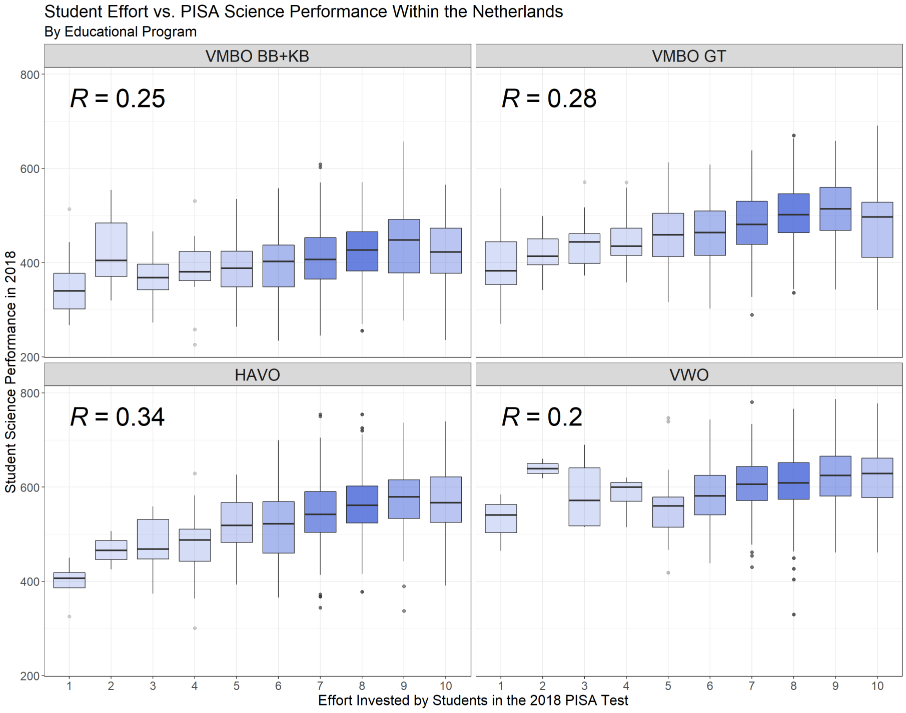


Figure 6.12: Correlation between effort and science ability by educational track

The VWO group showed the weakest correlation between effort and test performance among the investigated subpopulations (Reading: $R = 0.21$; Mathematics: $R = 0.15$; Science: $R = 0.20$). The distribution of effort scores was coherent across genders educational programs.

6.3.2 Summary

The self-reported effort of students in the Netherlands was moderately positively correlated with their estimated ability at the 2018 PISA cycle. The correlation was higher for female than male students and noticeably weaker for students in the VWO program. The findings were consistent across the reading, mathematics and science domains.

7. Discussion

The results of the Netherlands-specific analyses varied between the three investigated PISA domains. In the mathematics domain, half of the anchor items were flagged as functioning differently across PISA cycles. However the national trend in mathematics performance did not change when controlling for the country-specific DIF. This implies that the DIF detected in the mathematics items canceled out over the course of the assessment, and could thereby be attributed to random variation in the data collection process. The finding indicates that for the mathematics domain, a global calibration of anchor items is sufficient to capture possible item-mode and item-cycle interactions within the Netherlands.

In contrast, the analyses showed a systematic difference in the functioning of reading trend items between the 2015 and 2018 PISA cycles. Controlling for the systematic DIF led to an upward correction for the estimated reading literacy within the Netherlands at the 2018 cycle. A possible explanation is a Netherlands-specific item-cycle effect, which implies that a fixed item parameter linking approach between 2015 and 2018, as utilized in the PISA study, cannot be applied to the Netherlands.

An alternative explanation is a global effect of the change in test design on the performance of students, which would imply that the fixed item parameter linking approach is generally inappropriate to link PISA cycles that differ in the deployed test design. The matter becomes further complicated if the test design interacts with student-level variables such as motivation, commitment or test anxiety (e.g., Asseburg & Frey, 2013; Kimura, 2017; Ling et al., 2017; Martin & Lazendic, 2018). The student-level variables can differ between countries in their effect on performance, how they are affected by the test design, and in their average level across students within a country. This highlights the importance of carefully evaluating the impact of student-level variables on the PISA scores at each new cycle, which is in line with the recommendations of Zieger et al., 2020.

For the science domain, the global calibration of anchor items led to a systematic overestimation of science literacy within the Netherlands at the 2015 cycle. In the adjusted national trends, the drop in performance was equally pronounced for female and male students, but only observed for the VMBO educational groups. The findings imply a discrepancy in science trend item functioning between the Netherlands and the global population at the 2015 cycle. The discrepancy was not captured in the scaling procedure of the 2015 PISA study and thereby produced a biased estimate of the science literacy within the Netherlands.

Moreover, the science trend adjustment within the Netherlands was limited to the 2015 cycle and did not affect the performance estimates for 2012 and 2018. This is indicative of a Netherlands-specfic presentation mode effect that only occurred in first computer-based cycle. Given that the mode effect was limited to VMBO students and was not observable at the following computer-based cycle, it possibly reflects a onetime difficulty in the transitioning from paper-based to computer-based assessments within the Netherlands. The absence of a Netherlands-specific mode effect should be confirmed at the 2021 PISA cycle to ensure an efficient scaling procedure for the planned transition to the adaptive test design in 2024.

Finally, it was shown that the engagement of students during the test-taking process correlated positively with their test performance. On country-level, the correlation was weak for the reading domain and of moderate size for the mathematics and science domains. Within the Netherlands, the correlation was moderately positively for all domains and consistently higher for female than male students. The findings indicate that taking the students' test-taking effort into account when computing their plausible values can improve the accuracy of the PISA rankings as well as of the national trends. Given the inherent limitations of self-report measures, it is advised to include response-time based measures of effort when constructing an overall index of student test engagement for the PISA study (Michaelides et al., 2020; Pools & Monseur, 2021). Ideally such an index is invariant across countries and time and can thereby be related to national as well as global comparisons of test performance.

Bibliography

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists)*, *34*(5), 502–508. https://doi.org/10.1111/opo.12131

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit [Place: Germany Publisher: Pabst Science Publishers]. *Psychological Test and Assessment Modeling*, *55*(1), 92–104.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441. https://doi.org/10.1007/BF03173192

Butler, J., & Adams, R. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of applied measurement*, *8*, 279–304.

Finn, B. (2015). Measuring motivation in low-stakes assessments [Publisher: Wiley Online Library]. *ETS Research Report Series*, *2015*(2), 1–17.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring Success in Education: The Role of Effort on the Test Itself. *American Economic Review: Insights*, *1*(3), 291–308. https://doi.org/10.1257/aeri.20180633

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design [Publisher: Sage Publications Sage CA: Thousand Oaks, CA]. *Applied psychological measurement*, *26*(1), 3–24.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., Garnier-Villarreal, M., Selig, J., Boulton, A., Preacher, K., Coffman, D., Rhemtulla, M., Robitzsch, A., Enders, C., Arslan, R., Clinton, B., Panko, P., Merkle, E., Chesnut, S., . . . Johnson, A. R. (2021). semTools: Useful Tools for Structural Equation Modeling. Retrieved September 21, 2021, from https://CRAN.R-project.org/package=semTools

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*, *14*, 12. https://doi.org/10.3352/jeehp.2017.14.12

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior [Publisher: SAGE Publications Inc]. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., & Weib, M. (2002). German scale handbook for PISA 2000. *Berlin, Germany: Max-Planck-Institut für Bildungsforschung*.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? [Publisher: SAGE Publications Inc]. *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lumley, T. (2021). Survey: Analysis of Complex Survey Samples. Retrieved September 21, 2021, from https://CRAN.R-project.org/package=survey

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2021a). Dexter: Data Management and Analysis of Tests. Retrieved September 21, 2021, from https://CRAN.R-project.org/package=dexter

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2021b). dexterMML [original-date: 2020-09-29T14:14:46Z]. Retrieved November 2, 2021, from https://github.com/dexter-psychometrics/dexterMML

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings [Publisher: Routledge _eprint: https://doi.org/10.1207/s15328007sem1103_2]. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from Plausible Values? *Psychometrika*, *81*(2), 274–289. https://doi.org/10.1007/s11336-016-9497-x

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience [Place: US Publisher: American Psychological Association]. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items [Publisher: Routledge _eprint: https://doi.org/10.1080/15305058.2019.1706529].

International Journal of Testing, *20*(3), 187–205. https://doi.org/10.1080/15305058.2019.1706529

OECD. (2016). *PISA 2015 Technical Report* (tech. rep.). Organisation for Economic Co-operation and Development. Paris.

OECD. (2019a). *PISA 2018 Results (Volume I): What Students Know and Can Do*. Organisation for Economic Co-operation; Development. https://doi.org/10.1787/5f07c754-en

OECD. (2019b). *PISA 2018 Technical Report* (tech. rep.). Organisation for Economic Co-operation and Development. Paris.

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education*, *9*(1), 10. https://doi.org/10.1186/s40536-021-00104-6

Putnick, D. L., & Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental review : DR*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2021). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (2021). Lavaan: Latent Variable Analysis. Retrieved September 21, 2021, from https://CRAN.R-project.org/package=lavaan

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Steinfeld, J., & Robitzsch, A. (2021). Item Parameter Estimation in Multistage Designs: A Comparison of Different Estimation Approaches for the Rasch Model [Number: 3 Publisher: Multidisciplinary Digital Publishing Institute]. *Psych*, *3*(3), 279–307. https://doi.org/10.3390/psych3030022

Tijmstra, J., Bolsinova, M., Liaw, Y.-L., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for Detecting Item-Level Misfit in Low-Performing Countries [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jedm.12263]. *Journal of Educational Measurement*, *57*(4), 566–583. https://doi.org/10.1111/jedm.12263

Tong, X., & Bentler, P. M. (2013). Evaluation of a New Mean Scaled and Moment Adjusted Test Statistic for SEM [Publisher: Routledge _eprint: https://doi.org/10.1080/10705511.2013.742403]. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 148–156. https://doi.org/10.1080/10705511.2013.742403

Walker, C. M. (2011). What's the DIF? Why Differential Item Functioning Analyses Are an Important Part of Instrument Development and Validation [Publisher: SAGE Publications Inc]. *Journal of Psychoeducational Assessment*, *29*(4), 364–376. https://doi.org/10.1177/0734282911406666

Wise, S. L., & Cotten, M. R. (2009). Test-taking effort and score validity: The influence of student conceptions of assessment. *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187–205). Information Age Publishing.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions [Publisher: Routledge _eprint: https://doi.org/10.1207/s15326977ea1001_1]. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Gao, L. (2017). A General Approach to Measuring Test-Taking Effort on Computer-Based Tests [Publisher: Routledge _eprint: https://doi.org/10.1080/08957347.2017.1353992]. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Zieger, L., Jerrim, J., Anders, J., & Shure, N. (2020). *Conditioning: How background variables can influence PISA scores* (CEPEO Working Paper Series No. 20-09). Centre for Education Policy and Equalising Opportunities, UCL Institute of Education. Retrieved October 15, 2021, from https://econpapers.repec.org/paper/uclcepeow/20-09.htm

Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, *80*(1), 65–84. https://doi.org/10.1007/s11336-013-9369-6

A. Appendix

In Table A.1 the results of a DIF analysis between 2015 and 2018 can be found. Within this table statistical significant differences (DIF) between 2015 and 2018 in the parameters of anchor items for the reading domain within the Netherlands are reported, where the letter *A* refers to a difference in discrimination and *B* to a difference in difficulty. The same results for mathematics and science in Tables A.3 and A.5 respectively.

In Tables A.7, A.8 and A.9 the average ability for reading, mathematics and science within the Netherlands for the PISA population can be found. The estimates are based on different sets of anchor items that link the scales between years.

			DIF				
Item ID	**Aspect**	**Type (CB)**	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
R220Q01	Access and retrieve	Open Response - Computer Scored	A				
R420Q02	Access and retrieve	Open Response - Human Coded	A		B		
R420Q09	Access and retrieve	Open Response - Human Coded			B		
R446Q03	Access and retrieve	Complex Multiple Choice	B		B	B	
R453Q05	Access and retrieve	Complex Multiple Choice	B		B	B	
R456Q01	Access and retrieve	Simple Multiple Choice	B		B		B
R220Q04	Integrate and interpret	Simple Multiple Choice	B	B	B	B	B
R404Q03	Integrate and interpret	Simple Multiple Choice	B	B			
R404Q06	Integrate and interpret	Simple Multiple Choice	B		B		
R404Q07	Integrate and interpret	Complex Multiple Choice			B		
R406Q01	Integrate and interpret	Open Response - Human Coded	B		B		B
R406Q02	Integrate and interpret	Open Response - Human Coded	B	B	B		
R406Q05	Integrate and interpret	Open Response - Human Coded	B	B	B	B	B
R412Q06	Integrate and interpret	Complex Multiple Choice	A			B	B
R412Q08	Integrate and interpret	Open Response - Human Coded	B	A	B	B	
R420Q10	Integrate and interpret	Open Response - Human Coded	A				
R437Q01	Integrate and interpret	Simple Multiple Choice	B	B	B		B
R437Q07	Integrate and interpret	Open Response - Human Coded	A	A			
R453Q01	Integrate and interpret	Simple Multiple Choice	A	A	B	B	
R455Q04	Integrate and interpret	Simple Multiple Choice	AB			B	
R455Q05	Integrate and interpret	Complex Multiple Choice	A				B
R456Q02	Integrate and interpret	Open Response - Human Coded	B	B	B	B	
R456Q06	Integrate and interpret	Open Response - Human Coded	B		B		
R424Q03	Reflect and evaluate	Simple Multiple Choice	AB	A	B	B	
R446Q06	Reflect and evaluate	Open Response - Human Coded	B		B		
R453Q04	Reflect and evaluate	Open Response - Human Coded	B	B	B		B
R453Q06	Reflect and evaluate	Open Response - Human Coded	A				
R455Q02	Reflect and evaluate	Open Response - Human Coded	B	B	B	B	

Table A.1: Reading: DIF between 2015 and 2018

			DIF				
Item ID	**Aspect**	**Type (CB)**	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
R220Q01	Access and retrieve	Open Response - Computer Scored	B	AB	AB	AB	AB
R420Q02	Access and retrieve	Open Response - Human Coded	B	A	A	AB	A
R446Q03	Access and retrieve	Complex Multiple Choice	A		A	A	
R453Q05	Access and retrieve	Complex Multiple Choice	A		A		AB
R455Q03	Access and retrieve	Open Response - Human Coded	AB		AB	AB	
R456Q01	Access and retrieve	Simple Multiple Choice				B	
R466Q02	Access and retrieve	Open Response - Human Coded	B	B		B	B
R466Q06	Access and retrieve	Open Response - Computer Scored	AB		AB	B	
R220Q04	Integrate and interpret	Simple Multiple Choice	A	A	A		A
R404Q06	Integrate and interpret	Simple Multiple Choice				A	
R404Q07	Integrate and interpret	Complex Multiple Choice	B	B			B
R406Q01	Integrate and interpret	Open Response - Human Coded	A				A
R406Q02	Integrate and interpret	Open Response - Human Coded				AB	
R406Q05	Integrate and interpret	Open Response - Human Coded	A				A
R412Q06	Integrate and interpret	Complex Multiple Choice			B		A
R412Q08	Integrate and interpret	Open Response - Human Coded	A				B
R420Q10	Integrate and interpret	Open Response - Human Coded	B		B	B	
R424Q02	Integrate and interpret	Complex Multiple Choice	AB	A	AB		AB
R432Q01	Integrate and interpret	Open Response - Human Coded	A		AB	AB	
R437Q01	Integrate and interpret	Simple Multiple Choice					A
R437Q06	Integrate and interpret	Simple Multiple Choice	B			B	
R437Q07	Integrate and interpret	Open Response - Human Coded	B	B		B	
R453Q01	Integrate and interpret	Simple Multiple Choice					A
R455Q05	Integrate and interpret	Complex Multiple Choice	B		B		
R456Q02	Integrate and interpret	Open Response - Human Coded	A	A	A		
R456Q06	Integrate and interpret	Open Response - Human Coded	A	AB	A	B	
R420Q06	Reflect and evaluate	Open Response - Human Coded		B			
R424Q03	Reflect and evaluate	Simple Multiple Choice					B
R432Q05	Reflect and evaluate	Open Response - Human Coded	A				
R453Q04	Reflect and evaluate	Open Response - Human Coded	A				
R455Q02	Reflect and evaluate	Open Response - Human Coded					AB

Table A.2: Reading: DIF between PBA and CBA

			DIF				
Item ID	**Content**	**Type (CB)**	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
M909Q03	Change and Relationships	Open Response - Computer Scored	B		B	AB	
M915Q02	Change and Relationships	Open Response - Computer Scored	B				
M954Q01	Change and Relationships	Open Response - Computer Scored	B	B		B	
M954Q04	Change and Relationships	Open Response - Computer Scored	B				
M496Q02	Quantity	Open Response - Computer Scored				A	
M564Q01	Quantity	Simple Multiple Choice		B			
M906Q02	Quantity	Open Response - Human Coded	AB	AB	B	A	B
M909Q02	Quantity	Simple Multiple Choice	AB	B	A	B	
M034Q01	Space and Shape	Open Response - Computer Scored					A
M406Q02	Space and Shape	Open Response - Human Coded		B			
M949Q01	Space and Shape	Complex Multiple Choice	AB		AB		
M949Q03	Space and Shape	Open Response - Human Coded	B	B			
M992Q02	Space and Shape	Open Response - Computer Scored	AB				
M408Q01	Uncertainty and Data	Complex Multiple Choice					A
M423Q01	Uncertainty and Data	Simple Multiple Choice	A				A
M803Q01	Uncertainty and Data	Open Response - Computer Scored	B				
M953Q03	Uncertainty and Data	Open Response - Computer Scored	A		A		
M955Q03	Uncertainty and Data	Open Response - Computer Scored	AB	AB	B	B	AB
M982Q03	Uncertainty and Data	Complex Multiple Choice				A	A

Table A.3: Mathematics: DIF between 2015 and 2018

			DIF				
Item ID	**Content**	**Type (CB)**	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
M446Q02	Change and Relationships	Open Response - Human Coded				B	
M909Q03	Change and Relationships	Open Response - Computer Scored			A		
M943Q01	Change and Relationships	Simple Multiple Choice	B		B		
M954Q01	Change and Relationships	Open Response - Computer Scored			B		AB
M954Q02	Change and Relationships	Open Response - Human Coded	AB	B	AB	B	B
M954Q04	Change and Relationships	Open Response - Computer Scored		AB			B
M998Q04	Change and Relationships	Complex Multiple Choice			A	B	
M442Q02	Quantity	Complex Multiple Choice	B			AB	
M474Q01	Quantity	Simple Multiple Choice	B		B	B	B
M496Q02	Quantity	Open Response - Computer Scored	B				B
M559Q01	Quantity	Simple Multiple Choice	B		B		B
M564Q01	Quantity	Simple Multiple Choice	B		B		B
M603Q01	Quantity	Complex Multiple Choice	B	B	B	B	B
M828Q03	Quantity	Open Response - Computer Scored	B		B		B
M905Q02	Quantity	Open Response - Human Coded	B	B	B	B	B
M906Q01	Quantity	Simple Multiple Choice	B	B		B	
M909Q02	Quantity	Simple Multiple Choice		A	B		
M919Q02	Quantity	Open Response - Computer Scored	B	B			B
M00GQ01	Space and Shape	Open Response - Computer Scored	AB		AB		B
M00KQ02	Space and Shape	Open Response - Human Coded	AB	A			AB
M033Q01	Space and Shape	Simple Multiple Choice	AB	AB	AB	B	AB
M034Q01	Space and Shape	Open Response - Computer Scored	B	B	B		B
M273Q01	Space and Shape	Complex Multiple Choice	B	B	B	B	B
M305Q01	Space and Shape	Simple Multiple Choice	AB	AB	B	AB	B
M406Q01	Space and Shape	Open Response - Human Coded	AB	AB	AB		AB
M406Q02	Space and Shape	Open Response - Human Coded	AB	A	AB		AB
M943Q02	Space and Shape	Open Response - Computer Scored			A		
M949Q02	Space and Shape	Complex Multiple Choice		B			B
M992Q01	Space and Shape	Open Response - Computer Scored					B
M420Q01	Uncertainty and Data	Complex Multiple Choice				B	
M828Q02	Uncertainty and Data	Open Response - Human Coded				B	
M915Q01	Uncertainty and Data	Simple Multiple Choice					B
M953Q02	Uncertainty and Data	Open Response - Human Coded	B		B	B	B
M953Q03	Uncertainty and Data	Open Response - Computer Scored	B	AB		B	
M955Q01	Uncertainty and Data	Open Response - Human Coded				B	
M955Q03	Uncertainty and Data	Open Response - Computer Scored			A		
M982Q01	Uncertainty and Data	Open Response - Computer Scored	B		B		B
M982Q04	Uncertainty and Data	Simple Multiple Choice				B	

Table A.4: Mathematics: DIF between PBA and CBA

			DIF				
Item ID	**Competency**	**Type (CB)**	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
S415Q07	Evaluate and design scientific enquiry	Complex Multiple Choice	B				B
S415Q08	Evaluate and design scientific enquiry	Complex Multiple Choice	B	B	B	B	
S425Q04	Evaluate and design scientific enquiry	Open Response - Human Coded	B	B			
S438Q01	Evaluate and design scientific enquiry	Complex Multiple Choice		B			
S438Q02	Evaluate and design scientific enquiry	Simple Multiple Choice	B		B		
S438Q03	Evaluate and design scientific enquiry	Open Response - Human Coded	B	B	B	B	
S466Q01	Evaluate and design scientific enquiry	Complex Multiple Choice					B
S256Q01	Explain phenomena scientifically	Simple Multiple Choice	B			B	
S408Q01	Explain phenomena scientifically	Simple Multiple Choice	B		B		B
S408Q04	Explain phenomena scientifically	Complex Multiple Choice					B
S428Q05	Explain phenomena scientifically	Open Response - Human Coded	B	B	B	B	
S478Q03	Explain phenomena scientifically	Complex Multiple Choice	B		B	B	B
S514Q02	Explain phenomena scientifically	Open Response - Human Coded	B				
S527Q03	Explain phenomena scientifically	Complex Multiple Choice	B			B	
S413Q06	Interpret data and evidence scientifically	Complex Multiple Choice		A			
S428Q01	Interpret data and evidence scientifically	Simple Multiple Choice		B			
S478Q02	Interpret data and evidence scientifically	Complex Multiple Choice					B
S498Q04	Interpret data and evidence scientifically	Open Response - Human Coded	B		B		

Table A.5: Science: DIF between 2015 and 2018

			DIF				
Item ID	**Competency**	**Type (CB)**	*Pooled*	*Female*	*Male*	*VMBO*	*HAVO+VWO*
S408Q05	Evaluate and design scientific enquiry	Simple Multiple Choice	B			B	
S415Q07	Evaluate and design scientific enquiry	Complex Multiple Choice		B		B	
S425Q05	Evaluate and design scientific enquiry	Simple Multiple Choice		B			
S438Q01	Evaluate and design scientific enquiry	Complex Multiple Choice	AB		B	B	
S438Q02	Evaluate and design scientific enquiry	Simple Multiple Choice				B	
S498Q02	Evaluate and design scientific enquiry	Complex Multiple Choice			A		
S498Q03	Evaluate and design scientific enquiry	Simple Multiple Choice	AB	B	B	B	
S256Q01	Explain phenomena scientifically	Simple Multiple Choice	A				A
S326Q04	Explain phenomena scientifically	Complex Multiple Choice					A
S408Q03	Explain phenomena scientifically	Open Response - Human Coded		B			
S415Q02	Explain phenomena scientifically	Simple Multiple Choice	B			B	
S478Q03	Explain phenomena scientifically	Complex Multiple Choice	A	B			
S514Q02	Explain phenomena scientifically	Open Response - Human Coded	A	AB	B	B	AB
S514Q03	Explain phenomena scientifically	Open Response - Human Coded	B		B	B	
S326Q02	Interpret data and evidence scientifically	Open Response - Human Coded					A
S413Q04	Interpret data and evidence scientifically	Complex Multiple Choice	B	B	B	B	B
S413Q06	Interpret data and evidence scientifically	Complex Multiple Choice	A				
S425Q02	Interpret data and evidence scientifically	Simple Multiple Choice	B	B	B	B	B
S428Q01	Interpret data and evidence scientifically	Simple Multiple Choice	A	A			A
S428Q03	Interpret data and evidence scientifically	Simple Multiple Choice	A	A	A	B	
S514Q04	Interpret data and evidence scientifically	Open Response - Human Coded	A				
S527Q04	Interpret data and evidence scientifically	Complex Multiple Choice	B			B	

Table A.6: Science: DIF between PBA and CBA

Year	Group	Anchor Items	Est.	SE	CI.95.Lower	CI.95.Upper
2012	Pooled	PISA	512	1.6	508.9	515.2
2015	Pooled	PISA	507	1.7	503.8	510.4
2018	Pooled	PISA	491	1.3	487.9	493.1
2012	Pooled	No DIF (Pooled)	511	1.7	507.4	514.2
2015	Pooled	No DIF (Pooled)	507	1.7	503.7	510.6
2018	Pooled	No DIF (Pooled)	495	1.4	491.8	497.3
2012	Female	PISA	521	2.0	516.9	524.6
2015	Female	PISA	522	2.2	517.8	526.4
2018	Female	PISA	504	1.9	500.2	507.6
2012	Female	No DIF (Pooled)	519	2.2	514.4	523.0
2015	Female	No DIF (Pooled)	520	2.8	514.3	525.2
2018	Female	No DIF (Pooled)	507	2.4	502.4	511.9
2012	Female	No DIF (Female)	519	2.1	515.2	523.5
2015	Female	No DIF (Female)	521	2.9	515.6	526.9
2018	Female	No DIF (Female)	506	1.7	502.9	509.7
2012	Male	PISA	502	2.4	497.4	506.9
2015	Male	PISA	494	2.8	488.9	499.8
2018	Male	PISA	478	1.7	474.5	481.3
2012	Male	No DIF (Pooled)	501	2.5	496.3	506.0
2015	Male	No DIF (Pooled)	493	2.0	489.0	496.9
2018	Male	No DIF (Pooled)	481	2.1	477.2	485.3
2012	Male	No DIF (Male)	500	2.6	495.3	505.5
2015	Male	No DIF (Male)	495	2.9	489.1	500.5
2018	Male	No DIF (Male)	480	1.6	477.3	483.5
2012	VMBO BB+KB	PISA	407	2.9	401.2	412.8
2015	VMBO BB+KB	PISA	416	2.5	410.9	420.8
2018	VMBO BB+KB	PISA	403	1.8	399.3	406.3
2012	VMBO BB+KB	No DIF (Pooled)	406	2.6	401.2	411.4
2015	VMBO BB+KB	No DIF (Pooled)	414	3.3	407.1	420.1
2018	VMBO BB+KB	No DIF (Pooled)	405	2.0	400.7	408.7
2012	VMBO BB+KB	No DIF (VMBO)	407	2.3	403.0	411.9
2015	VMBO BB+KB	No DIF (VMBO)	414	3.8	406.7	421.8
2018	VMBO BB+KB	No DIF (VMBO)	404	2.4	399.8	409.0
2012	VMBO GT	PISA	477	3.0	471.1	482.9
2015	VMBO GT	PISA	475	3.2	468.8	481.4
2018	VMBO GT	PISA	464	2.5	459.1	469.0
2012	VMBO GT	No DIF (Pooled)	476	2.7	470.3	480.7
2015	VMBO GT	No DIF (Pooled)	472	2.9	466.8	478.2
2018	VMBO GT	No DIF (Pooled)	465	2.8	459.6	470.5
2012	VMBO GT	No DIF (VMBO)	476	2.2	471.3	479.9
2015	VMBO GT	No DIF (VMBO)	474	2.8	468.0	479.1
2018	VMBO GT	No DIF (VMBO)	464	2.8	458.3	469.2
2012	HAVO	PISA	540	2.7	535.0	545.5
2015	HAVO	PISA	546	3.5	538.6	552.4
2018	HAVO	PISA	536	2.6	530.7	540.8
2012	HAVO	No DIF (Pooled)	539	2.2	535.1	543.7
2015	HAVO	No DIF (Pooled)	546	2.4	541.0	550.4
2018	HAVO	No DIF (Pooled)	536	2.2	531.5	540.0
2012	HAVO	No DIF (HAVO+VWO)	540	2.9	534.3	545.8
2015	HAVO	No DIF (HAVO+VWO)	545	2.5	540.1	550.0
2018	HAVO	No DIF (HAVO+VWO)	537	2.1	533.0	541.4
2012	VWO	PISA	602	2.0	598.2	606.0
2015	VWO	PISA	600	4.3	591.6	608.6
2018	VWO	PISA	589	2.5	584.3	594.0
2012	VWO	No DIF (Pooled)	601	2.2	596.8	605.6
2015	VWO	No DIF (Pooled)	598	3.7	590.8	605.3
2018	VWO	No DIF (Pooled)	591	2.2	586.6	595.2
2012	VWO	No DIF (HAVO+VWO)	602	2.5	597.2	607.1
2015	VWO	No DIF (HAVO+VWO)	598	2.8	592.6	603.6
2018	VWO	No DIF (HAVO+VWO)	591	3.6	583.7	597.6

Table A.7: Trends in reading ability in the Netherlands

Year	Group	Anchor Items	Est.	SE	CI.95.Lower	CI.95.Upper
2012	Pooled	PISA	529	1.8	525.5	532.5
2015	Pooled	PISA	514	1.2	511.9	516.8
2018	Pooled	PISA	521	1.2	518.7	523.3
2012	Pooled	No DIF (Pooled)	529	1.6	525.8	532.1
2015	Pooled	No DIF (Pooled)	515	1.3	512.3	517.5
2018	Pooled	No DIF (Pooled)	522	1.3	519.8	524.7
2012	Female	PISA	523	2.1	519.1	527.5
2015	Female	PISA	511	1.7	507.2	513.9
2018	Female	PISA	520	1.9	516.1	523.6
2012	Female	No DIF (Pooled)	523	2.3	518.9	527.8
2015	Female	No DIF (Pooled)	512	1.7	509.0	515.6
2018	Female	No DIF (Pooled)	520	1.6	516.9	523.1
2012	Female	No DIF (Female)	524	2.2	519.3	528.0
2015	Female	No DIF (Female)	511	2.1	507.1	515.3
2018	Female	No DIF (Female)	519	1.6	515.7	522.2
2012	Male	PISA	534	1.8	530.9	537.9
2015	Male	PISA	516	1.8	512.6	519.5
2018	Male	PISA	524	2.2	519.3	528.0
2012	Male	No DIF (Pooled)	535	1.7	531.4	538.2
2015	Male	No DIF (Pooled)	517	1.8	513.4	520.5
2018	Male	No DIF (Pooled)	524	2.0	520.0	527.8
2012	Male	No DIF (Male)	534	2.2	529.9	538.6
2015	Male	No DIF (Male)	517	1.9	513.2	520.8
2018	Male	No DIF (Male)	524	1.9	520.1	527.4
2012	VMBO BB+KB	PISA	431	2.2	426.8	435.4
2015	VMBO BB+KB	PISA	429	2.0	424.7	432.6
2018	VMBO BB+KB	PISA	428	2.1	424.1	432.4
2012	VMBO BB+KB	No DIF (Pooled)	431	1.9	427.0	434.4
2015	VMBO BB+KB	No DIF (Pooled)	428	2.3	423.2	432.2
2018	VMBO BB+KB	No DIF (Pooled)	428	1.9	424.6	432.2
2012	VMBO BB+KB	No DIF (VMBO)	430	2.1	426.3	434.5
2015	VMBO BB+KB	No DIF (VMBO)	428	2.4	423.4	432.8
2018	VMBO BB+KB	No DIF (VMBO)	428	2.2	423.8	432.5
2012	VMBO GT	PISA	501	1.9	497.2	504.7
2015	VMBO GT	PISA	496	2.1	491.7	500.0
2018	VMBO GT	PISA	497	2.2	492.4	500.9
2012	VMBO GT	No DIF (Pooled)	501	1.7	497.1	503.9
2015	VMBO GT	No DIF (Pooled)	494	2.0	490.6	498.3
2018	VMBO GT	No DIF (Pooled)	497	2.1	493.1	501.3
2012	VMBO GT	No DIF (VMBO)	500	1.7	497.0	503.6
2015	VMBO GT	No DIF (VMBO)	495	2.1	491.1	499.3
2018	VMBO GT	No DIF (VMBO)	498	2.6	492.5	502.9
2012	HAVO	PISA	560	1.8	557.0	563.9
2015	HAVO	PISA	559	1.8	555.7	562.6
2018	HAVO	PISA	559	2.0	555.0	563.0
2012	HAVO	No DIF (Pooled)	561	1.6	557.9	564.0
2015	HAVO	No DIF (Pooled)	560	1.9	555.8	563.3
2018	HAVO	No DIF (Pooled)	559	2.1	555.0	563.2
2012	HAVO	No DIF (HAVO+VWO)	561	1.8	557.5	564.6
2015	HAVO	No DIF (HAVO+VWO)	560	1.4	557.1	562.6
2018	HAVO	No DIF (HAVO+VWO)	559	2.4	554.4	563.7
2012	VWO	PISA	617	1.8	613.5	620.5
2015	VWO	PISA	606	2.2	601.2	610.0
2018	VWO	PISA	611	2.4	606.8	616.0
2012	VWO	No DIF (Pooled)	615	1.8	611.2	618.1
2015	VWO	No DIF (Pooled)	608	2.6	603.1	613.3
2018	VWO	No DIF (Pooled)	611	2.6	606.3	616.6
2012	VWO	No DIF (HAVO+VWO)	616	2.0	611.7	619.5
2015	VWO	No DIF (HAVO+VWO)	608	2.5	602.7	612.4
2018	VWO	No DIF (HAVO+VWO)	612	2.4	607.0	616.5

Table A.8: Trends in mathematics ability in the Netherlands

Year	Group	Anchor Items	Est.	SE	CI.95.Lower	CI.95.Upper
2012	Pooled	PISA	525	1.8	521.7	528.7
2015	Pooled	PISA	511	1.8	507.4	514.6
2018	Pooled	PISA	502	1.6	499.3	505.7
2012	Pooled	No DIF (Pooled)	528	2.3	523.2	532.1
2015	Pooled	No DIF (Pooled)	504	2.1	499.7	508.1
2018	Pooled	No DIF (Pooled)	505	1.5	501.5	507.6
2012	Female	PISA	522	2.4	517.6	527.2
2015	Female	PISA	510	2.7	505.2	515.6
2018	Female	PISA	505	2.5	499.8	509.5
2012	Female	No DIF (Pooled)	526	2.6	520.9	531.0
2015	Female	No DIF (Pooled)	503	2.8	497.5	508.7
2018	Female	No DIF (Pooled)	505	2.0	501.2	508.9
2012	Female	No DIF (Female)	523	2.3	518.8	527.6
2015	Female	No DIF (Female)	504	2.4	499.0	508.5
2018	Female	No DIF (Female)	505	2.3	500.6	509.7
2012	Male	PISA	528	2.6	522.7	532.8
2015	Male	PISA	511	3.1	505.5	517.5
2018	Male	PISA	502	2.5	497.2	506.8
2012	Male	No DIF (Pooled)	530	2.6	524.5	534.8
2015	Male	No DIF (Pooled)	505	3.7	497.7	512.4
2018	Male	No DIF (Pooled)	503	2.1	499.3	507.6
2012	Male	No DIF (Male)	529	2.4	524.3	533.9
2015	Male	No DIF (Male)	506	3.0	499.7	511.4
2018	Male	No DIF (Male)	504	2.4	499.3	508.9
2012	VMBO BB+KB	PISA	424	2.5	419.4	429.3
2015	VMBO BB+KB	PISA	412	3.2	405.6	418.2
2018	VMBO BB+KB	PISA	404	2.0	400.5	408.5
2012	VMBO BB+KB	No DIF (Pooled)	427	2.7	421.9	432.4
2015	VMBO BB+KB	No DIF (Pooled)	406	2.5	400.6	410.6
2018	VMBO BB+KB	No DIF (Pooled)	407	3.0	400.6	412.4
2012	VMBO BB+KB	No DIF (VMBO)	426	3.0	419.9	431.6
2015	VMBO BB+KB	No DIF (VMBO)	406	3.2	399.3	412.1
2018	VMBO BB+KB	No DIF (VMBO)	406	2.1	401.9	410.1
2012	VMBO GT	PISA	496	2.1	491.9	500.2
2015	VMBO GT	PISA	489	3.2	482.4	494.8
2018	VMBO GT	PISA	484	2.0	480.5	488.4
2012	VMBO GT	No DIF (Pooled)	498	2.7	493.0	503.4
2015	VMBO GT	No DIF (Pooled)	484	2.4	479.6	489.1
2018	VMBO GT	No DIF (Pooled)	485	2.7	480.0	490.6
2012	VMBO GT	No DIF (VMBO)	499	2.6	494.4	504.6
2015	VMBO GT	No DIF (VMBO)	484	2.6	479.3	489.4
2018	VMBO GT	No DIF (VMBO)	486	2.0	482.1	489.8
2012	HAVO	PISA	557	2.5	552.2	561.9
2015	HAVO	PISA	556	3.1	550.2	562.3
2018	HAVO	PISA	552	2.7	546.9	557.7
2012	HAVO	No DIF (Pooled)	558	2.8	552.7	563.6
2015	HAVO	No DIF (Pooled)	553	3.0	547.0	558.9
2018	HAVO	No DIF (Pooled)	553	2.5	547.7	557.6
2012	HAVO	No DIF (HAVO+VWO)	557	2.1	552.8	560.9
2015	HAVO	No DIF (HAVO+VWO)	556	2.8	550.1	561.0
2018	HAVO	No DIF (HAVO+VWO)	551	2.9	545.8	557.0
2012	VWO	PISA	615	2.8	609.2	620.0
2015	VWO	PISA	609	3.2	602.9	615.4
2018	VWO	PISA	608	2.9	602.4	613.7
2012	VWO	No DIF (Pooled)	616	2.7	611.1	621.7
2015	VWO	No DIF (Pooled)	604	3.4	597.2	610.6
2018	VWO	No DIF (Pooled)	607	2.7	601.3	612.0
2012	VWO	No DIF (HAVO+VWO)	615	2.4	610.2	619.6
2015	VWO	No DIF (HAVO+VWO)	608	3.0	602.2	614.1
2018	VWO	No DIF (HAVO+VWO)	608	2.9	601.8	613.2

Table A.9: Trends in science ability in the Netherlands

List of Figures

List of Tables