Test Performance during PISA Assessments

Within the Netherlands and across OECD countries

Stichting Cito

Onderzoek, Kennis & Innovatie

Auteurs
Konrad Klotzke
Remco Feskens

1. Executive Summary

The students' test-taking performance can vary during low-stake assessments, such as the PISA survey. Comparisons between PISA cycles and/or across countries can be problematic if the magnitude of decline differs between the compared groups. In this report, we explore the consistency in decline across OECD countries and between different sub-populations within the Netherlands (gender, educational program, socio-economic status, immigration status). Mathematics, reading and science data from the 2012, 2015 and 2018 PISA assessments are analyzed.

Performance decline across OECD countries

- **Development of test performance:** The development of test performance over the course of the PISA assessment differed substantially between PISA cycles and across OECD countries. In 2012, a monotone decline in performance between subsequent cluster positions was observed for the mathematics, reading and science domains. At the 2015 and 2018 cycles, a different pattern in the performance trend was observed: students performed worse after they were presented a 30-minutes cluster of items measuring the same domain, compared to two clusters of items covering a different domain. The pattern was most pronounced at the science domain and in low-performing countries. It did moreover consistently occur in combination with the first pattern, i.e., the average performance across clusters that measure the same domain was lower when the block of clusters was placed at the end of a booklet. The results are indicative of cluster order effects that possibly interact with cognitive factors related to the students' invested effort during the test-taking process (e.g., boredom, task enjoyment, self-control), and can thereby substantially distort comparisons between PISA cycles and across countries.
- **Test performance and country ability:** At several domain-cycle combinations, low-performing countries tended to exhibit a stronger score variation between cluster

positions than countries placed at the upper half of the corresponding domain-cycle performance ranking. The finding can be partly attributed to a stronger correlation between the students' performance level and their persistence during PISA assessments in low-performing countries. The country-level performance did however not comprehensively explain the differences in cluster score variation that were observed across the investigated PISA cycles and countries.

Performance decline within the Netherlands

- **Gender:** The Netherlands-specific analyses revealed no substantial differences between male and female students in the development of test performance over the course of the PISA assessments. The most noticeable difference between gender groups was found in the proportion responses coded as non-response or *not-reached* at the end of the assessments, which tended to be higher for female students. The difference was most pronounced at the 2018 science domain. A possible explanation can be found in the higher level of motivation exhibited by female students at low-stake assessments, which could be linked to an earlier onset of rapid-guessing behaviour of male students in the PISA assessments.
- **Educational programs and socio-economic status:** Moderate overlap was found between educational programs and socio-economic status (ESCS) groups in the performance development and for the patterns of missing response data over the course the PISA assessments. A plausible explanation can be found in a confounding role of the students' proficiency level, which is likely correlated with both the followed educational program and the ESCS index. The results moreover provided further support for the role of the students' proficiency level in predicting the development of their test performance over the course of the PISA assessments. However, similar to the OECD country-level analyses, the measured performance could only explain part of the variation in order effects between subpopulations of Dutch students.
- **Ethnicity:** Students native to the Netherlands showed a consistently higher performance and less score variation across cluster positions than first- and second-generation students. The finding was moreover reflected by the proportions responses coded as non-response or *not-reached* that remained more stable for the native students over the course of the PISA assessments. Due to the small size of the first- and second generation groups, the results should however be interpreted with caution.

Contents

2. Introduction

The Program for International Student Assessment (PISA) is designed to measure the performance of 15-year-old students across several cognitive domains and countries (OECD, 2019a). A key premise of the PISA study is the comparability of performance estimates across assessments taken in different years, countries, and between groups within a country (e.g., genders or educational programs). Naturally, the premise of comparability is threatened by the possibility of systematic interactions between the measured test performance and student- or group-level factors.

In this report, we investigate if the development of the students' performance during PISA assessments varies between members of the Organisation for Economic Co-operation and Development (OECD) and in a second step, between subpopulations within the Netherlands. In general, it is expected that the test-taking performance declines over the course of a low-stake assessment (List et al., 2017; M. Wu, 2010), i.e., the probability of a correct response for equally proficient students is lower if the item is placed further near the end of the assessment. The PISA study does not explicitly control for the decline in performance and thereby implicitly assumes that the shifts in performance are homogeneous across the compared groups. Conversely, if the assumption of homogeneity is violated, the group-specific performance measurements differ in the extend to which they are affected by factors that contribute to the decline in performance over the course of the assessment.

Factors that can, directly or indirectly, affect the performance during a low-stake assessment include the effort that students invest into the assessment (Asseburg & Frey, 2013; Finn, 2015; Wise & Cotten, 2009), their task enjoyment (M. A. Lindner et al., 2019; Penk et al., 2014), test anxiety (Ling et al., 2017) and physical or mental fatigue (C. Lindner et al., 2018; Wise & Gao, 2017). Systematic group-level differences in the level of performance-altering factors can consequently distort the comparison between countries, PISA cycles or subpopulations (e.g., male and female students). As a remedy, the PISA study utilizes various self-reported measures related to the students' test-taking effort,

enjoyment and anxiety in the process of obtaining performance estimates. However, due to their subjective nature self-report measures must be interpreted with caution (Kong et al., 2007; Wise & Gao, 2017). Furthermore, their inclusion in the statistical model of the PISA study was criticised for its lack of transparency and the unclear effect on the PISA country scores (Zieger et al., 2020).

In the PISA study, the group-level performance can be followed across four subsequent 30-minute clusters of items. In the 2015 and 2018 assessments, the majority of students were presented two blocks of items. Each block consists of two clusters and measures the students' performance at either the mathematics, reading or science domain. The standard booklets in the 2012 assessment were constructed slightly differently, with each booklet containing a minimum of one, and a maximum of three clusters for the main domain (mathematics). Most importantly, the items were rotated evenly across the four cluster positions of the PISA assessments, whereby responses to all items are available at each cluster position. Given the rotated cluster design, the performance, the non-response and the proportion responses coded as *not-reached* in a group can be monitored over the course of the assessments. The obtained patterns lend insight into group-level differences in the expected performance decline, and thereby indicate the degree to which comparisons between groups are affected by factors that determine the development of the students' performance during the test-taking process.

The report is structured as follows: chapter 3 outlines the data analyzed and the methodology applied in this study. In chapter 4 the interaction between the development of the students' performance during PISA assessments and the country in which they were presented the assessment is investigated. Moreover, we explore possible interactions between the gender, followed education program, socio-economic status and immigration status of Dutch students with their observed pattern in performance during the 2012, 2015 and 2018 PISA assessments. Chapter 5 summarizes the results and provides advice for future research. Finally, Appendix A contains additional results for the Netherlands-specific analyses.

3. Method

Response data from the 2012, 2015 and 2018 PISA studies are analyzed. The data were collected in members of the Organisation for Economic Co-operation and Development (OECD) and encompass responses from the three PISA domains reading, mathematics and science. The list of participating OECD members comprised 34 countries in 2012, 35 in 2015 (New: Latvia) and 37 in 2018 (New: Colombia and Lithuania). In 2012 all responses stem from paper-based assessments (PBA), in 2015 and 2018 all students were presented the computer-based assessment (CBA) form. Across PISA cycles, for each country the response data from the standard booklets and, if applicable, their easier variants are selected. In the 2012 assessment design, the booklets 21-27 contain the easier mathematics cluster variants but are otherwise equal to the standard booklets 1-7 (Figure 3.1; PR: Reading, PM: Mathematics, PS: Science cluster). In 2015 and 2018, both the normal and easier cluster variants are part of the standard booklets (Figure 3.2 and 3.3; R: Reading, M: Mathematics, S: Science cluster).

The chosen set of booklets represents a balanced design in which each cluster is equally rotated across the four cluster positions. Thereby it can be investigated if the order of the clusters affects the students' test performance and if the performance diminishes over the course of the assessment. Moreover, the booklet design allows the inclusion of response data from low-performing countries that chose to administer easier cluster variants. The special one-hour ("Une Heure", UH) booklets differ both in length and cluster composition from the standard booklets and are excluded from the analysis.

In 2015 the rotation of the science clusters followed a randomized multistep process. The exact position of a student's response in the booklet design can be identified through lookup tables that link the allocated base test form and a random number to the cluster location (OECD, 2016, Chapter 2). A multistage adaptive testing (MSAT) design was deployed in 2018 for the reading domain. In contrast to the randomized cluster rotation in 2015, the location of responses in the MSAT design does not translate to the examined

cluster positions in the booklet design. The data of the 2018 reading assessment are therefore not used in this study.

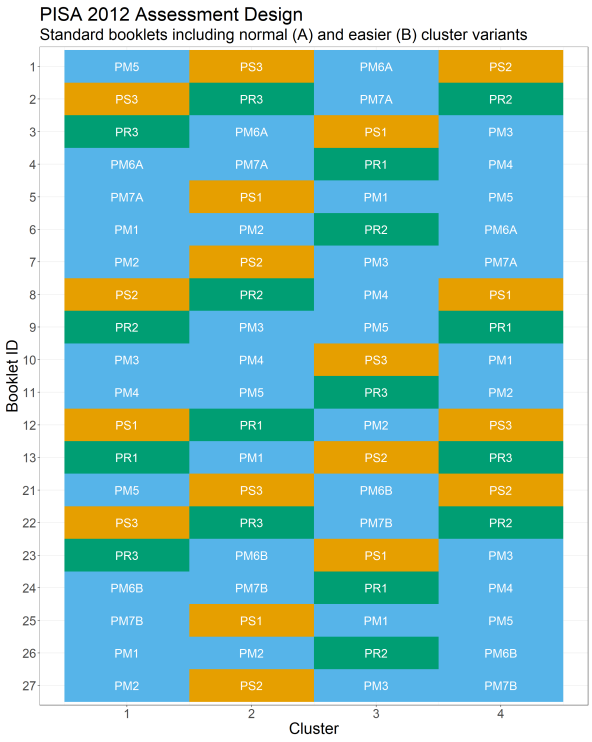

Figure 3.1: PISA 2012 assessment design.

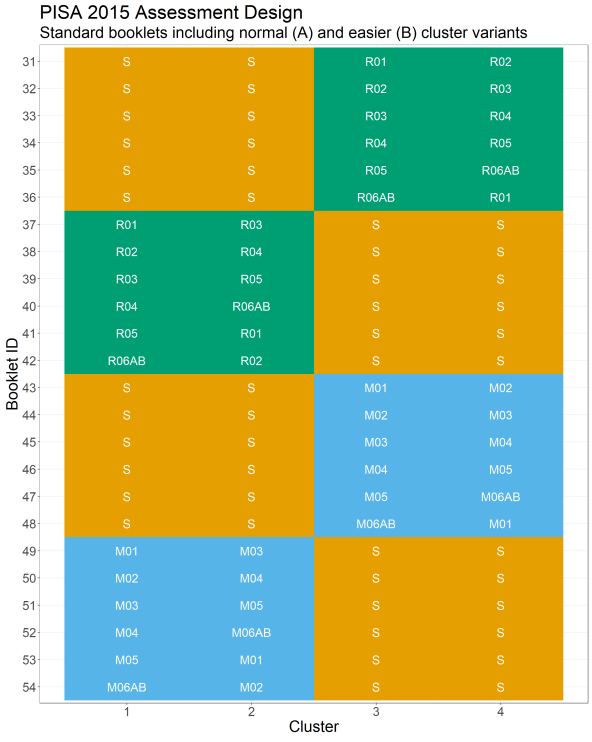

Figure 3.2: PISA 2015 assessment design.

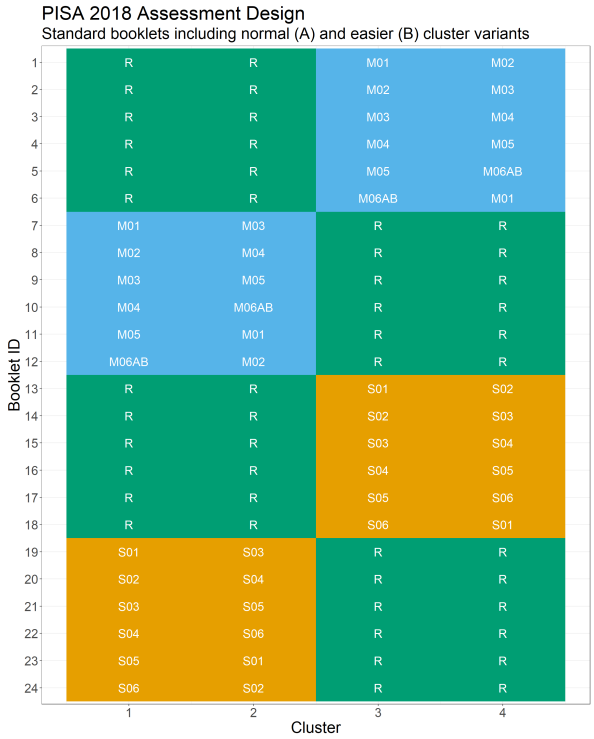

Figure 3.3: PISA 2018 assessment design.

The Netherlands-specific analyses are split by the student-level variables gender, educational program, the PISA index of economic, social and cultural status (ESCS) and the PISA index of immigration status (IMMIG). The group-specific sample sizes are shown in Table 3.1, 3.2 and 3.3.

Educational programs for which a program-year group contains responses from less than 50 students are excluded from the data. The PISA ESCS index is calculated based on the parents' highest level of education and occupation status, and household possessions such as a personal computer, books or educational software. For each year, students are split into five ordinal categories where the highest category identifies the 20% of students for which the highest ESCS was recorded. The PISA IMMIG index allocates students to one of three categories: (1) students who had at least one parent born in the country where the assessment took place (native students), (2) students who were born in the county of assessment and whose parents were born in a different country (second-generation students), and (3) students who, like their parents, were born outside of the country of assessment (first-generation students). Further information about the PISA indices can be found in OECD, 2019b, Annex A1.

Across the three investigated PISA cycles only little data are available from first-generation students in the Netherlands. Results from the group of first-generation students are therefore prone to be strongly affected by measurement error and must be interpreted with caution. The word of caution also applies to inferences about the VMBO BB group that took the PISA assessment in 2018. The small sample of VMBO BB students in 2018 compared to previous cycles can be attributed to a higher proportion of students that took the UH form of the assessment. Overall, 17.9% of the assessed Dutch students were presented the UH form instead of a standard booklet in 2018, compared to 2.7% in 2015 and 2.9% in

2012. The corresponding numbers for the VMBO BB group are 71.7% (2018), 0.5% (2015) and 0% (2012). Similarly, the proportion of VMBO KB (2018: 34.2%; 2015: 0.1%; 2012: 0%), first-generation (2018: 48.3%; 2015: 8.8%; 2012: 6.8%) and second-generation students (2018: 29.5%; 2015: 4.0%; 2012: 4.2%) that took the UH form and are therefore not included in the analysed data increased in 2018.

	Group	**2012**	**2015**	**2018**
Gender	Female	2030	886	840
	Male	2122	806	877
Educational program	VMBO BB	377	138	58
	VMBO KB	531	243	192
	VMBO GT	1159	492	521
	HAVO	1079	460	492
	VWO	1006	359	454
ESCS (quantile)	1	831	339	344
	2	831	339	344
	3	830	338	343
	4	830	338	343
	5	830	338	343
Immigration status	Native	3713	1511	1508
	Second-generation	324	119	166
	First-generation	89	25	23

Table 3.1: Number of students in the Netherlands at the mathematics domain.

	Group	**2012**	**2015**
Gender	Female	1386	827
	Male	1461	832
Educational program	VMBO BB	267	131
	VMBO KB	359	237
	VMBO GT	797	473
	HAVO	748	428
	VWO	676	390
ESCS (quantile)	1	570	332
	2	570	332
	3	569	332
	4	569	332
	5	569	331
Immigration status	Native	2547	1466
	Second-generation	222	137
	First-generation	58	30

Table 3.2: Number of students in the Netherlands at the reading domain.

	Group	2012	2015	2018
Gender	Female	1402	1713	900
	Male	1463	1640	868
Educational program	VMBO BB	260	269	50
	VMBO KB	381	480	225
	VMBO GT	786	966	521
	HAVO	751	888	514
	VWO	687	750	458
ESCS (quantile)	1	573	671	354
	2	573	671	354
	3	573	671	354
	4	573	670	353
	5	573	670	353
Immigration status	Native	2558	2979	1571
	Second-generation	228	256	143
	First-generation	62	55	34

Table 3.3: Number of students in the Netherlands at the science domain.

The performance of students is evaluated by computing the group-average percentage of the correct responses across all items in the investigated domain. Responses to items were scored as 0 (incorrect) or 1 (correct), or, if applicable, as 0 (incorrect), 1 (partial credit), or 2 (full credit). The computation of the percentage correct responses consists of three steps. First of all, all responses are divided by the maximum score category of the corresponding items. Thereby the dichotomous items remain scored as 0 or 1, and the categories of the polytomously scored items are scaled to 0, 0.5 and 1, where 0.5 represents a partial credit. Secondly, the arithmetic mean across all responses is computed and thirdly, the resulting value is multiplied by 100 to obtain the percentage correct responses across all items and students that are part of the investigated group and domain. The procedure is summarized in Equation 3.1:

$$p = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} Y_{ik}/C_k}{NK} \cdot 100, \tag{3.1}$$

where N and K represent the number of students in the group and the number of items in the domain, respectively. The maximum score category for item k is indicated by C_k.

Furthermore, the group-average percentages for two types of missing response data are computed. Item-level non-response is defined as a missing value that was recorded when a student was expected to provide a response to an item. A series of successive missing values at the end of a cluster are coded as *not-reached*, excluding the first missing value in the series which is coded as non-response. For the 2012 PISA cycle, missing data are excluded from the computation of the percentage correct responses, while for 2015 and 2018 the missing responses are coded as incorrect and therefore affect the students' measured test performance.

To quantify the severity of differences in the performance across cluster positions, the

root mean squared deviation (RMSD) is computed as follows for each country:

$$RMSD = \sqrt{\frac{\sum_{c=1}^{4} (p_c - \bar{p})^2}{4}}, \tag{3.2}$$

where p_c and $\bar{p}$ are the cluster position-specific and year-average percentage correct responses in a country. The RMSD therefore represents the average deviation in percentage correct responses from the year-average across cluster positions.

4. Results

4.1 Mathematics

In Figure 4.1, 4.2 and 4.3 the percentage correct responses to mathematics items of students in OECD countries is shown for the 2012, 2015 and 2018 PISA cycles. The numbers are computed per cluster position to provide insight into the development of the students' test performance over the course of the PISA assessment. The figures furthermore contain the OECD-average percentages for item-level non-response and for response data coded as *not-reached*.

In 2012, 2015 and 2018 the OECD-average mathematics performance gradually decreased over the course of the PISA assessment. The trend was consistent across countries at the 2012 and 2015 cycles, however in 2018 a subset of OECD countries showed a higher performance in the third compared to the second mathematics cluster position. The finding contradicts the hypothesis of a gradually decreasing test performance over time and was predominantly observed for low-performing countries (Colombia, Mexico, Chile, Greece, Israel), which can be indicative of country-level factors that interact with the clusters' position in the test form, or generally affect the response- and/or test-taking process. A notable exception to the low-performing countries formed Sweden, which saw a strong decrease of performance in the second mathematics cluster of a booklet (1: 48.6%; 2: 43.2%; 3: 46.2%; 4: 41.4%). As shown by the regression model applied in Figure 4.5, in 2018 a change of 1 point in the percentage correct responses on average corresponds to a change of 5.3 points in the PISA mathematics score. For Sweden the drop in performance between the first and last cluster position therefore can be translated to a decline in PISA score of $(48.6 - 41.4) \cdot 5.3 = 38.2$ points. The relationship between performance and PISA score illustrated in Figure 4.5 moreover indicates that for 2018, the Netherlands scored lower than expected on the official scale given the average percentage correct responses in the country. The derivation can be explained by the high (relative to previous cycles)

proportion of Dutch students that were presented the UH test form. The special UH form was mainly deployed for selected students in low-performing groups (e.g., the VMBO BB and VMBO KB educational programs). Its exclusion from this study thereby naturally excludes a set of low-performing students from the computation of the percentage correct responses. Regardless, across OECD countries the correlation between the percentage correct responses and the PISA score was very high at all three cycles (2012: .95; 2015: .96; 2018: .97).

Figure 4.4 contains a quantification of the severity of the variation in performance between cluster positions. The numbers represent the average deviation of the position-specific percentages correct responses from the country-average percentage, the latter being, as discussed earlier, strongly correlated with the officially reported PISA score. As expected, Sweden saw a strong average deviation of 2.8 percent points across cluster positions, which translates to a RMSD in PISA score of 11.8. Further strong deviations were found in 2012 for Greece (average 3.2 %) and Mexico (average 2.5%), which implies that for those countries the PISA score would noticeably differ if it was computed on a subset of cluster positions, and is therefore indicative of an interaction between cluster position and the students' test performance. In contrast, the lowest deviations from the country-average were found in 2018 for Poland and Hungary (both .2%), which were therefore not affected by the position of the mathematics clusters in the presented test forms. The OECD-average deviation saw a drop from 1.6% in 2012 to 1.0% in 2015 and remained stable in 2018 (1.1%). For the Netherlands a constant decline in RMSD across PISA cycles was observed (2012: 2.1%; 2015: 1.9%; 2018: 1.6%), i.e., the differences in test performance between cluster positions became less severe at later PISA cycles. At the three investigated PISA cycles, the OECD-average percentage of item-level non-responses gradually increased over the course of the assessment. In 2012. the percentage item-level *not-reached* saw a slight increase at the third cluster position (1: .1%; 2: .2%; 3: 1.3%) and a noticeable jump at the fourth position (5.7%). In 2015 (1: .8%; 2: 1.5%; 3: 1.1%; 4: 1.4%) and 2018 (1: 1.1%; 2: 4.5%; 3: 1.7%; 4: 3.2%) the percentage *not-reached* was elevated at the second and fourth position. The following analysis provides further insight into patterns of missingness within the Netherlands at the PISA mathematics domain.

Percentage Correct Responses at the PISA 2012 Mathematics Domain
By Cluster Position

	1	2	3	4
South Korea	60.6	59.5	57.8	57.2
Japan	58.6	57.3	54.8	54.7
Belgium	54.4	54.8	52.8	51.5
Netherlands*	54.8	54.5	51.9	49.7
Czechia	53.8	53.2	52.8	51.1
Estonia	53.9	53.9	52.2	50.6
Switzerland	54.2	54.0	51.8	50.4
Poland	54.3	53.8	51.0	48.9
Germany	52.7	52.6	51.2	49.8
Canada	51.9	51.5	49.7	48.4
Finland	51.3	51.1	50.0	48.6
Austria	51.2	51.1	49.4	48.7
France	50.1	50.0	47.7	46.9
New Zealand	49.5	50.0	47.1	47.1
Ireland	48.9	49.3	47.4	46.5
Spain	49.6	49.1	46.7	44.7
Iceland	50.0	48.5	46.0	44.5
Australia	48.3	48.4	45.9	45.5
Italy	49.3	48.0	46.0	44.3
Norway	49.0	47.8	45.5	44.5
Luxembourg	47.9	47.6	45.4	43.8
United Kingdom	47.6	47.1	45.1	43.3
Slovakia	47.9	46.0	45.1	42.7
Denmark	46.7	46.5	45.0	43.5
Slovenia	47.5	46.9	44.5	42.6
Portugal	47.8	46.8	45.0	41.9
Sweden	47.5	45.6	44.0	42.0
Hungary	46.2	45.7	44.3	42.7
United States	45.1	45.5	42.9	42.1
Israel	45.5	43.9	42.5	39.2
Greece	42.6	40.9	37.2	34.3
Chile	40.3	39.1	37.7	35.7
Turkey	37.6	37.8	36.2	35.0
Mexico	35.9	34.3	32.1	29.3
OECD average	**48.1**	**47.4**	**45.5**	**44.0**
No Response	**8.3**	**9.2**	**11.0**	**11.5**
Not Reached	**0.1**	**0.2**	**1.3**	**5.7**

Figure 4.1: Performance at the 2012 mathematics domain by cluster position.

Percentage Correct Responses at the PISA 2015 Mathematics Domain
By Cluster Position

	1	2	3	4
Japan	53.8	54.4	54.6	54.0
South Korea	53.3	53.0	51.5	51.8
Estonia	52.6	51.9	50.7	50.0
Belgium	51.7	51.8	48.9	49.1
Netherlands*	52.0	51.7	49.2	47.5
Switzerland	52.4	51.2	48.4	47.4
Germany	49.5	50.2	49.4	49.4
Finland	49.1	48.6	49.4	48.9
Czechia	49.6	48.7	46.7	46.4
Canada	48.5	47.9	47.1	46.7
Denmark	49.5	47.9	45.7	45.9
Ireland	46.4	47.9	46.9	47.2
Poland	46.5	47.3	47.3	47.0
New Zealand	48.2	47.7	46.1	45.9
France	47.9	47.8	46.5	45.6
Italy	48.4	46.0	47.0	45.5
Austria	46.9	48.0	45.9	45.4
Norway	47.7	47.5	45.8	44.2
Slovenia	48.3	46.8	44.8	44.5
United Kingdom	46.8	47.2	44.2	43.5
Spain	46.2	45.1	44.7	43.1
Sweden	47.1	45.2	43.6	41.0
Luxembourg	45.1	44.9	43.5	43.0
Australia	45.0	45.0	43.2	42.1
Hungary	44.5	44.0	43.3	43.0
Iceland	44.3	43.0	44.4	42.8
Portugal	44.4	42.9	42.4	41.4
Latvia	43.6	43.2	41.5	41.8
Slovakia	44.4	41.4	41.9	41.3
United States	42.1	42.3	40.6	40.7
Israel	44.1	41.5	41.0	38.8
Greece	41.4	39.4	38.6	37.2
Chile	38.6	36.9	34.8	34.2
Mexico	32.7	31.0	30.8	29.6
Turkey	31.7	31.2	29.8	29.4
OECD average	**46.6**	**46.0**	**44.7**	**44.0**
No Response	**4.9**	**5.9**	**6.1**	**6.9**
Not Reached	**0.8**	**1.5**	**1.1**	**1.4**

Figure 4.2: Performance at the 2015 mathematics domain by cluster position.

Percentage Correct Responses at the PISA 2018 Mathematics Domain
By Cluster Position

	1	2	3	4
Netherlands*	54.8	54.8	51.3	51.8
South Korea	53.2	52.4	51.6	51.7
Japan	53.4	51.9	51.7	51.7
Estonia	52.8	51.6	50.0	49.6
Poland	50.5	50.1	50.1	49.9
Czechia	52.2	50.1	49.4	48.2
Belgium	51.2	50.5	48.8	48.8
Switzerland	51.0	48.7	48.1	47.2
Germany	49.0	48.8	47.2	47.4
Canada	49.6	48.3	47.1	47.0
Finland	48.8	47.2	47.4	45.5
Austria	47.5	47.3	46.1	46.4
Slovenia	48.8	47.7	45.5	45.2
Denmark	48.5	46.7	45.5	44.7
New Zealand	47.7	46.6	45.7	45.3
Ireland	46.1	47.0	45.6	46.0
Norway	48.6	45.5	45.5	43.5
United Kingdom	47.3	46.6	44.4	44.1
Italy	47.0	45.0	45.3	43.9
Portugal	47.2	44.8	44.2	43.9
Australia	47.2	46.1	43.5	43.2
Sweden	48.6	43.2	46.2	41.4
Hungary	44.0	44.5	44.6	44.3
Iceland	47.5	43.4	43.7	42.5
Latvia	44.9	44.5	43.6	43.9
France	45.7	43.8	43.9	42.9
Spain	45.8	44.2	43.2	42.8
Slovakia	45.7	43.5	43.4	43.1
Luxembourg	44.6	43.6	42.2	42.1
Lithuania	42.4	42.0	41.2	42.2
United States	43.1	41.9	40.8	39.9
Israel	42.0	38.2	39.7	37.4
Turkey	39.0	38.4	37.5	37.1
Greece	38.6	36.2	37.6	35.1
Chile	37.6	33.8	34.4	33.1
Mexico	32.5	27.0	30.7	28.5
Colombia	30.9	26.0	26.8	25.4
OECD average	**46.5**	**44.9**	**44.2**	**43.5**
No Response	**5.0**	**5.7**	**6.4**	**6.9**
Not Reached	**1.1**	**4.5**	**1.7**	**3.2**

Figure 4.3: Performance at the 2018 mathematics domain by cluster position.

RMSD Percentage Correct Responses at the PISA Mathematics Domain
Across Cluster Positions

	2012	2015	2018
South Korea	1.3	0.8	0.6
Japan	1.7	0.3	0.7
Netherlands*	2.1	1.9	1.6
Estonia	1.4	1.0	1.3
Belgium	1.3	1.4	1.1
Switzerland	1.6	2.0	1.4
Czechia	1.0	1.3	1.5
Germany	1.2	0.3	0.8
Poland	2.2	0.3	0.2
Finland	1.1	0.3	1.2
Canada	1.4	0.7	1.1
Austria	1.1	1.0	0.6
New Zealand	1.3	1.0	0.9
Ireland	1.1	0.5	0.5
France	1.4	1.0	1.0
Denmark	1.3	1.6	1.4
Italy	1.9	1.1	1.1
Norway	1.8	1.4	1.8
Slovenia	2.0	1.5	1.5
United Kingdom	1.7	1.6	1.4
Spain	2.0	1.1	1.2
Australia	1.3	1.2	1.7
Iceland	2.1	0.7	1.9
Sweden	2.0	2.2	2.8
Luxembourg	1.7	0.9	1.0
Portugal	2.2	1.1	1.3
Hungary	1.4	0.6	0.2
Slovakia	1.9	1.3	1.0
Latvia	Not OECD	0.9	0.5
United States	1.4	0.8	1.2
Lithuania	Not OECD	Not OECD	0.5
Israel	2.3	1.9	1.8
Greece	3.2	1.5	1.3
Chile	1.7	1.7	1.7
Turkey	1.1	1.0	0.7
Mexico	2.5	1.1	2.1
Colombia	Not OECD	Not OECD	2.2
OECD average	**1.6**	**1.0**	**1.1**
No Response	**1.3**	**0.7**	**0.7**
Not Reached	**2.3**	**0.3**	**1.3**

Figure 4.4: RMSD at the mathematics domain across cluster positions.

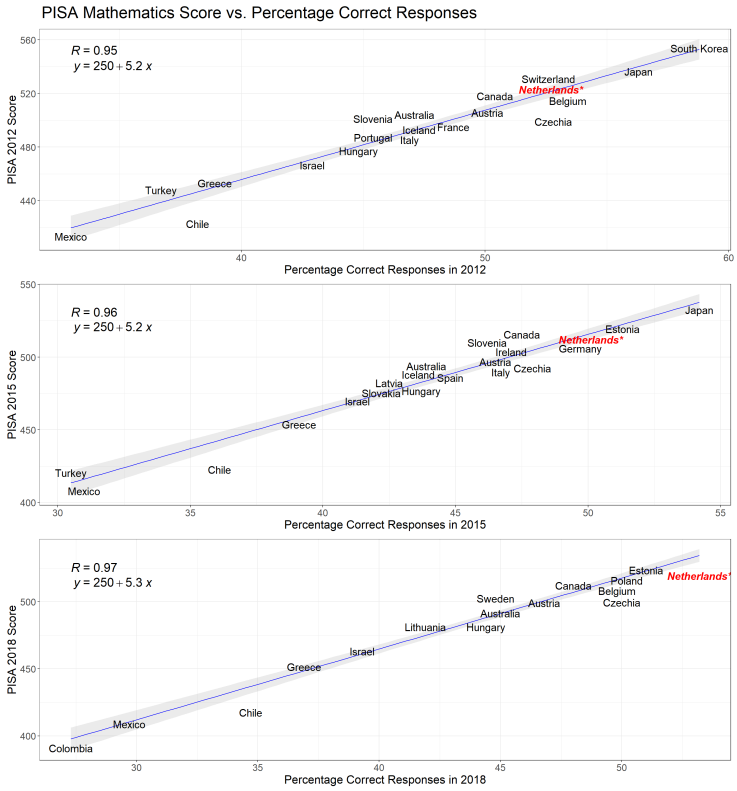

Figure 4.5: PISA mathematics score vs. percentage correct responses.

Figure 4.6, 4.7 and 4.8 illustrate developments in the percentage correct responses to mathematics items, the item-level percentage of non-response, and the percentage *not-reached* over the course of the PISA assessments for students in the Netherlands. The corresponding numbers can be found in the Appendix (Table A.1, A.2 and A.3).

Female students in the Netherlands showed a noticeable drop in the percentage correct responses at the third cluster position at all three PISA cycles. In 2012 and 2015, their performance declined further for mathematics clusters that were placed at the last position of a booklet. However in 2018 the opposite trend was observed, with the performance of female students seeing a slight increase at the fourth compared to the third position. In 2012, both genders saw a similar trend in performance over the course of the assessment, whereby male students achieved a higher percentage correct responses at all four cluster positions.

In contrast to female students, at the 2015 cycle male students already saw a decline in the percentage correct responses at the second cluster position and thereby a monotone degradation in mathematics performance over the course of the assessment. As a result, male students noticeably outperformed female students at the first cluster position (male: 53.5%; female: 51.6%) and dropped to an equal level of performance at the last position (male: 48.1%; female: 48.2%). For the 2018 cycle, the results did not indicate a difference in the development of mathematics performance between male and female students.

In 2012 and 2018, the level of non-response steadily increased at subsequent cluster positions for both genders. In the 2015 assessment data both genders saw a raise in non-response between the first and second cluster position. However for male students, the

non-response remained on an equal level between the second and third cluster position, and saw a steep incline at the fourth position. In contrast, the level of non-response for female students increased between the second and third cluster position and slightly declined at the fourth position. For both genders, a constant percentage *not-reached* was observed at the first three cluster positions of the 2012 assessment. At the fourth position, the percentage increased equally for both groups. In 2015, male and female students saw a steady level of *not-reached* at the first three cluster positions. At the last position the level of *not-reached* increased for the female students and decreased for the male students.

At the 2018 cycle, an equal level of *not-reached* was observed between genders at the first and third cluster position. At the second and fourth position the female group showed an elevated level of *not-reached* compared to the male group. It must be noted however, that the magnitude of the differences between cluster positions in the percentage *not-reached* was generally small at all cycles and for both genders. The largest discrepancy between cluster positions was found for the female group between the first and fourth position in 2015 (1: .5%; 4: 1.3%) and in 2018 (1: .5%; 4: 1.3%).

In 2012, the percentage correct responses of the VWO group was on an equal level across the first mathematics clusters in a booklet (cluster position 1 and 3), however it saw an incline between position 1 and 2 and a decline between position 3 and 4. HAVO students showed a constant performance at the first two cluster positions and a steady decline at the subsequent positions. For the VMBO groups a monotone decline in performance was observed over the course of the 2012 assessment.

At the 2015 assessment, the performance of VWO students increased at the second cluster position and declined to a constant level at positions 3 and 4. The performance of HAVO students did not noticeably differ between cluster position 1 and 2 and saw a slight decline at position 3 and 4. The VMBO GT and VMBO BB groups saw a steady decline in the percentage correct responses over the course of the assessment, with the strongest drops in performance occurring at the respective second mathematics cluster in a booklet (cluster position 2 or 4). For VMBO KB students an increase in performance at the second cluster position was observed, followed by a steep decline at the following positions.

In 2018, the performance of VWO students was elevated at cluster position 2 and 4. For the HAVO and VMBO GT groups a steep decline in performance was observed at the third cluster position. The performance recovered at the fourth cluster position, but remained below the level observed at position 1 and 2. The VMBO KB group saw a steady decline in mathematics performance over the course of the 2018 assessment. The trend in the VMBO BB group saw an elevation at cluster position 3 but was overall negative. Note that the results for the 2018 VMBO BB group are subject to strong random variation due to the small sample size.

The level of non-response differed widely between educational programs. In general, low-performing groups showed a stronger increase and a higher overall level of non-response at all three PISA cycles. The most notable difference between cycles was observed for the VMBO KB and HAVO groups. The VMBO KB group saw a steep incline in the level of non-response at varying cluster positions across cycles (2012: position 4; 2015: position 2; 2018: position 2 and 3). In the HAVO group, the level of non-response steadily increased over the course of the 2012 assessment. In 2015 the non-response increased at the second cluster position and remained stable at the subsequent positions. In 2018 a steep incline in

non-response was observed for HAVO students at the third and fourth cluster position.

In the 2012 assessment data the level of *not-reached* was generally low across educational programs. Notable exceptions were observed for the VMBO BB and VMBO GT groups, which saw an incline in the percentage *not-reached* at the fourth cluster position (VMBO BB: 1.4%; VMBO GT: 1.2%). In 2015 and 2018 the VMBO BB and VMBO KB groups produced the most remarkable results. At the 2015 cycle, VMBO KB students showed a steep incline in the level of *not-reached* at the third and fourth cluster position. In contrast, in the 2018 assessment data a decline at position 3 and 4 was observed. In both 2015 and 2018, an elevated level of *not-reached* was found for the VMBO BB group at the second cluster position. The elevation was most strongly pronounced in 2018. Interestingly, with the exception of the elevation found at the second cluster position, the level of *not-reached* decreased for VMBO BB students over the course of the 2015 and 2018 assessments. The results for the 2018 VMBO BB group must be however interpreted with caution given the small size of the program-year group.

The analysis of the ESCS groups indicated that students with a higher ESCS systematically performed better at the mathematics domain in all three PISA cycles. Most noticeable, for the highest ESCS quantile, the development of performance across the 2015 and 2018 assessments resembled the corresponding patterns of the VWO group. The level of non-response was on average higher at lower ESCS quantiles, however no systematic relationship between ESCS and the non-response pattern across cluster positions was observed. Similarly, the ESCS did not systematically affected the pattern *not-reached* across cluster positions. In 2012, all quantiles saw a incline in the percentage *not-reached* at the last cluster position. In 2015 the percentage was on average higher for lower ESCS quantiles. However at the 2018 assessment, students in the highest quantile showed the the highest level of *not-reached*.

Finally, students native to the Netherlands showed a higher mathematics performance than first- and second-generation students across all cluster positions and all three PISA cycles. In 2012 and 2015, the percentage correct responses of native students did not noticeably change between the first and second cluster position, and steadily declined at the subsequent positions. At the 2018 cycle, the performance of native students dropped noticeably between the second and third cluster position. Moreover, their performance was slightly higher at the second mathematics cluster in a booklet (cluster position 2 or 4) compared to the respective first cluster (cluster position 1 or 3).

First- and second-generation students showed an equal level of performance at the first cluster position of the 2012 assessment. Both groups saw a decrease in performance over the course of the assessment, however the decline was stronger for the second-generation students. The performance of the second-generation group at the 2015 assessment did not noticeable change between the first and second cluster position and decreased at the subsequent positions. In 2018 their performance saw steep drops at the second and fourth position. At the 2015 and 2018 cycles, the performance of first-generation students varied strongly between cluster positions. Due to the small sample size of the 2015 and 2018 first-generation groups, the variation must be interpreted with caution and does not provide clear evidence in favour of a systematic difference in mathematics performance between cluster positions.

Across PISA cycles, the second-generation group showed showed a stronger increase

and a higher overall level of non-response compared to the native students. In 2012, the first-generation students showed the highest level of non-response among the investigated immigration groups. The results at the 2015 and 2018 cycles were inconclusive due to the small sample of the first-generation student population.

In the 2012 assessment data, the level of *not-reached* was close to zero for all the three immigration groups. At the fourth cluster position, native and first-generation students showed a slight increase in the percentage *not-reached*. In 2015 and 2018, a higher level of *not-reached* was observed across all cluster positions for the second-generation compared to the native students. Most noticeably, in 2018 the second-generation group saw a steady increase in the percentage *not-reached* between the first and third cluster position, and a decline at the last position. For the native students a different pattern was observed: the level of *not-reached* increased between cluster position 1 and 2, dropped at the third position and remained steady at the last position.

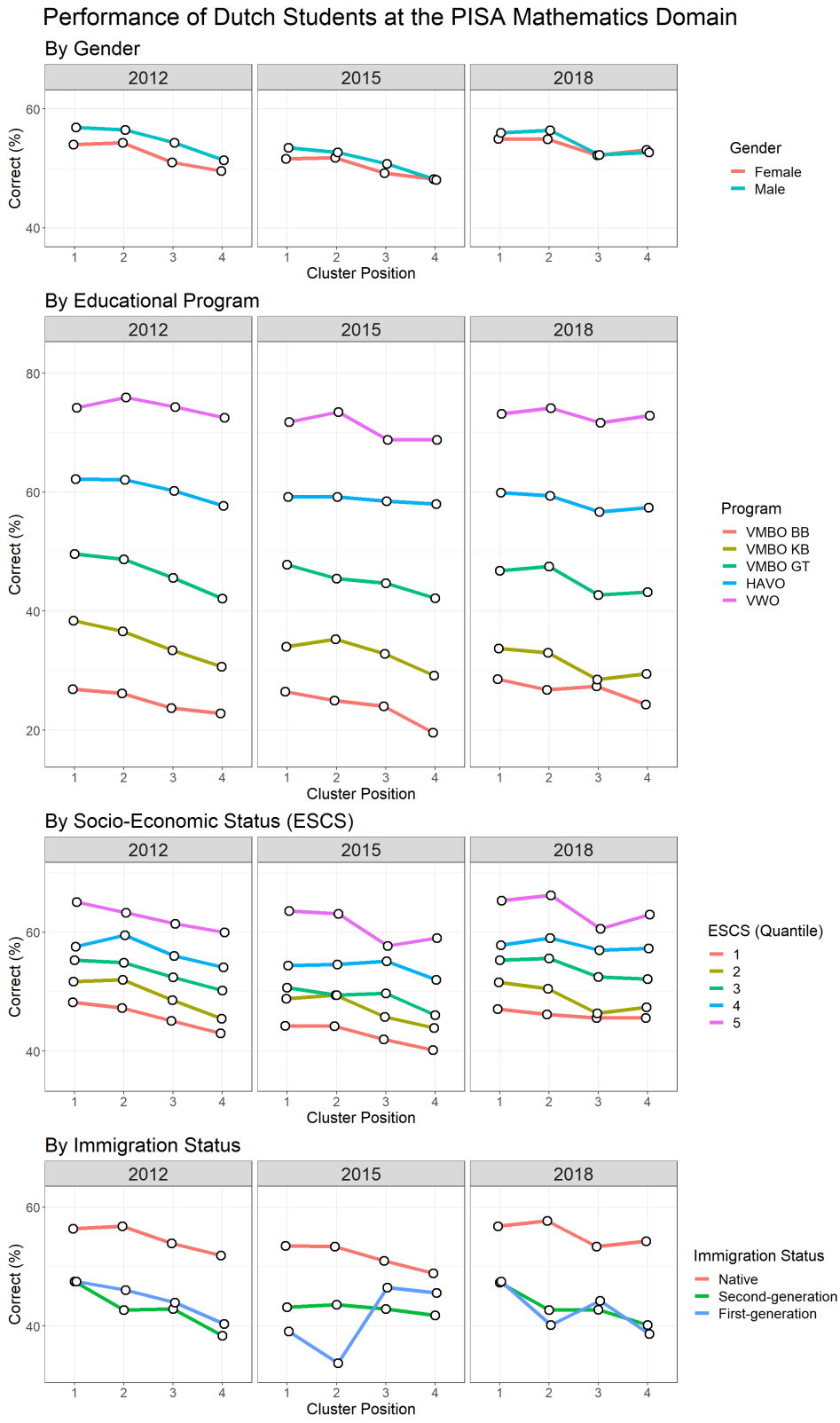

Figure 4.6: Performance of Dutch students at the PISA mathematics domain.

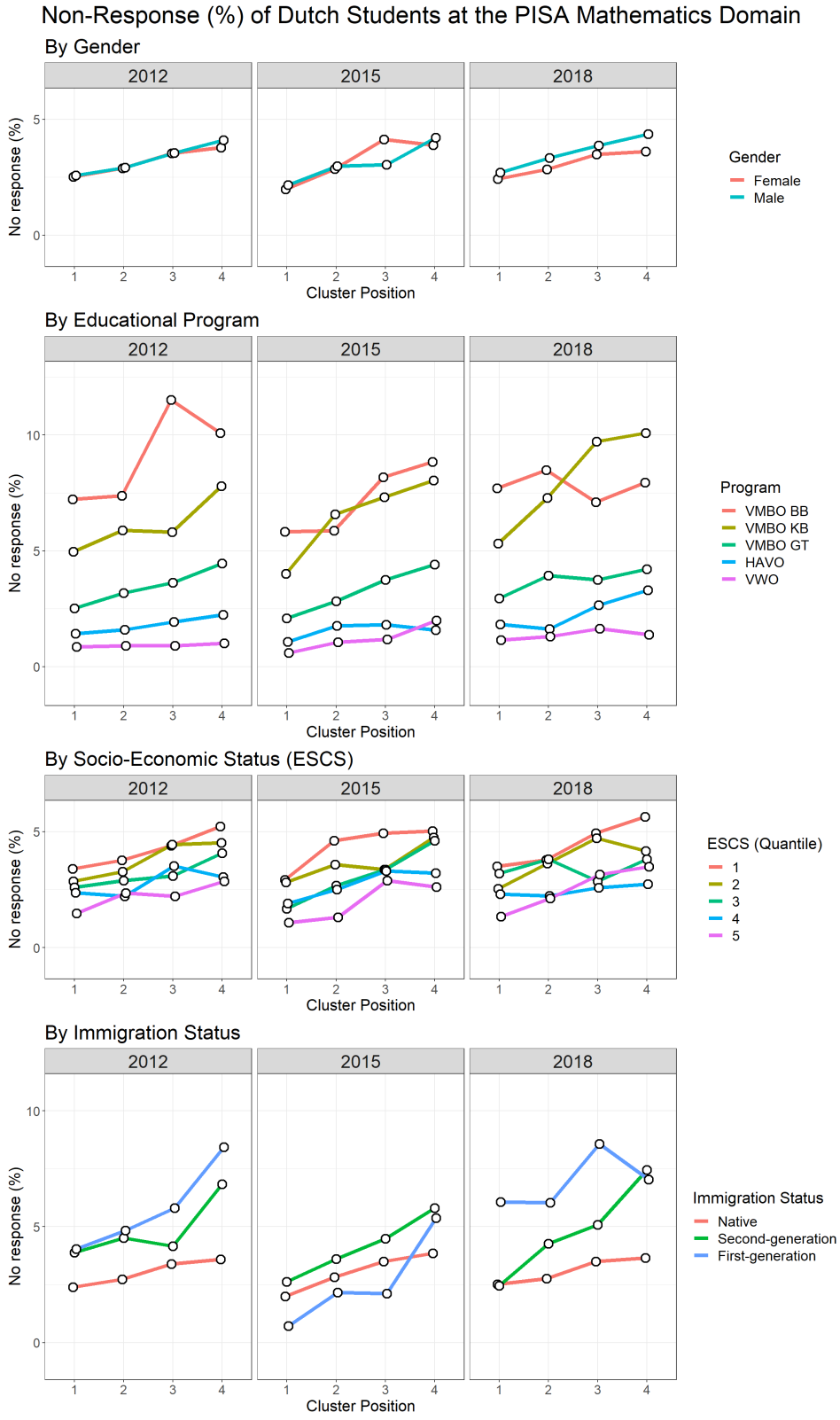

Figure 4.7: Non-Response (%) of Dutch students at the PISA mathematics domain.

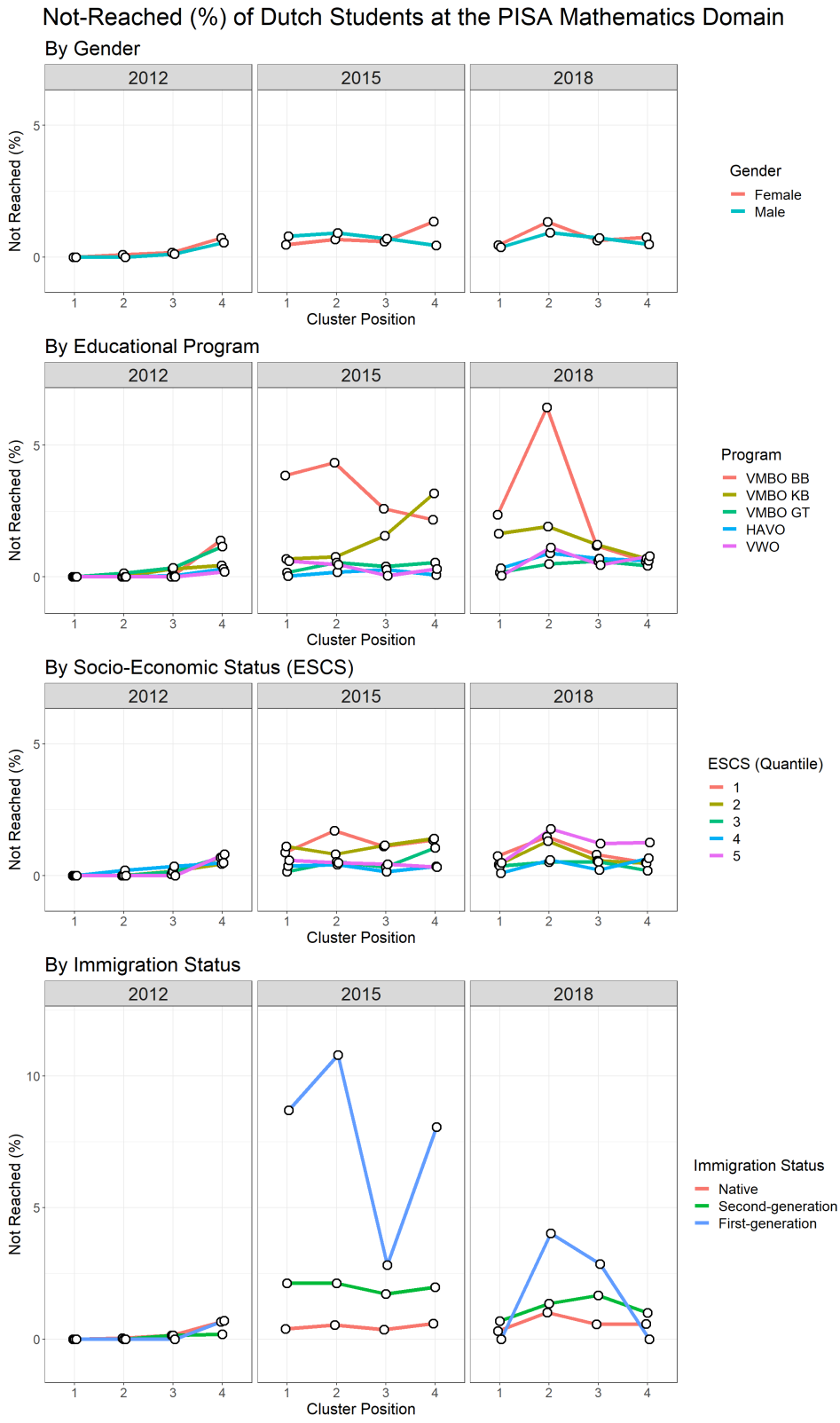

Figure 4.8: Not-Reached (%) of Dutch students at the PISA mathematics domain.

4.2 Reading

In Figure 4.9 and 4.10 the percentage correct responses to reading items of students in OECD countries is shown for the 2012 and 2015 PISA cycles. The numbers are computed per cluster position to provide insight into the development of the students' test performance over the course of the PISA assessment. The figures furthermore contain the OECD-average percentages for item-level non-response and for response data coded as *not-reached*.

In 2012 the OECD-average reading performance gradually decreased over the course of the PISA assessment. With the exception of Japan, which showed an inclined performance between the second and third cluster position (2: 63.8%; 3: 64.9%), the trend was consistent across countries at the 2012 cycle. At the 2015 assessment, the average performance across OECD countries declined at the second and fourth cluster position, but remained stable between position 2 and 3. The OECD-average was a mixture of three patterns in the country-level numbers: one subset of the OECD countries showed a steady decline in reading performance over the course of the 2015 assessment, the second set did not see a change between the second and third position and hence matched the pattern found in the OECD-average, and the third set saw an increased performance in the third compared to the second position. The first pattern was most pronounced in Belgium (2: 63%; 3: 60.3%) and the Netherlands (2: 64.4%; 3: 62.9%). The second pattern was found for countries across the entire performance spectrum and the third pattern was most prominent in the low-performing countries Greece (2: 52.2%; 3: 56.1%), Slovakia (2: 51.5%; 3: 53.3%) and Israel (2: 55.3%; 3: 57.1%). The mixed findings at the 2015 cycle are indicative of country-level factors that interact with the clusters' position in the test form, or generally affect the response- and/or test-taking process.

Figure 4.11 contains a quantification of the severity of the variation in performance between cluster positions. The numbers represent the average deviation of the position-specific percentages correct responses to reading items from the country-average percentage at the reading domain. On average, the variation in performance between cluster positions in OECD countries was greater in 2012 (3%) compared to 2015 (2.4%). The country-level percentages can be translated to a more meaningful scale by applying a regression model to the relationship between the percentage correct responses and the PISA reading score. As shown in Figure 4.12, in 2012 a change of 1 point in the percentage correct responses on average corresponds to a change of 5 points in the PISA reading score. The regression slope was slightly steeper at the 2015 assessment (5.4 points). The translated numbers show that the average variation in PISA score across cluster positions was greater in 2012 ($3 \cdot 5 = 15$ points) than at the 2015 cycle ($2.4 \cdot 5.4 = 13$ points).

At the 2012 assessment, countries with a higher variation in performance between cluster positions could be found predominately at the lower half of the performance spectrum. Noticeable exceptions were Poland (3.9%), the Netherlands (3.3%), New Zealand (3.3%) and Japan (3.1%), for which a RMSD in PISA score of 15 or greater was observed. In 2015, seven of the ten best performing countries showed a variation in performance between cluster positions that was noticeably smaller than the variation in the lower-performing countries. Exceptions were the Netherlands (2.5%), Belgium (2.6%) and Canada (2.4%). For the Netherlands, the RMSD in PISA score declined from $3.3 \cdot 5 = 16.5$ points in 2012 to $2.5 \cdot 5.4 = 13.5$ points in 2015, whereby the measured

reading performance depended less on the cluster position at the 2015 cycle.

Figure 4.12 moreover shows a high country-level correlation between the percentage correct responses to reading items and the PISA score at the reading domain. The correlation was higher in 2012 (.94) than at the 2015 cycle (.90), which is indicative of a stronger consistency in the translation of the 2012 percentages to the PISA scoring scale.

The OECD-average percentage of item-level non-responses gradually increased over the course of the 2012 and 2015 assessments. In 2012. the percentage item-level *not-reached* saw a slight increase at the third cluster position (1: .1%; 2: .2%; 3: 1.1%) and a steep incline at the fourth position (7.1%). At the 2015 cycle, the percentage *not-reached* was elevated at the second and fourth position (1: .3%; 2: 1.2%; 3: .4%; 4: .9%). The following analysis provides further insight into patterns of missingness within the Netherlands at the PISA reading domain.

Percentage Correct Responses at the PISA 2012 Reading Domain
By Cluster Position

	1	2	3	4
Japan	71.2	63.8	64.9	63.5
South Korea	67.6	63.2	63.1	62.1
Ireland	65.4	62.8	62.5	60.2
Poland	66.4	64.7	61.3	56.3
Estonia	65.9	63.6	60.0	58.9
Belgium	65.2	62.1	61.1	59.0
New Zealand	66.4	61.0	59.7	57.6
Canada	65.0	61.4	60.2	57.9
Czechia	64.7	61.0	59.6	58.1
France	65.0	60.6	60.0	57.2
Germany	63.1	62.0	58.6	56.2
Netherlands*	64.6	61.0	58.8	55.5
Finland	62.9	59.5	59.2	57.7
Norway	62.7	59.7	59.9	55.0
Australia	62.6	59.2	57.5	54.6
United States	62.5	57.8	57.5	55.0
United Kingdom	61.1	57.9	57.2	54.8
Israel	62.5	60.1	56.0	51.1
Switzerland	61.6	57.6	57.0	53.3
Italy	62.4	58.3	55.9	51.9
Hungary	60.0	57.7	56.7	52.3
Austria	58.8	57.2	56.8	53.6
Spain	60.8	56.6	55.9	51.3
Sweden	60.2	56.6	55.7	52.0
Luxembourg	60.2	56.6	55.0	51.3
Denmark	58.8	55.0	55.1	52.7
Portugal	60.7	55.9	53.6	49.8
Greece	60.8	55.4	53.3	44.6
Iceland	58.9	53.0	52.9	49.1
Turkey	57.2	55.1	53.6	47.8
Slovenia	55.9	51.9	49.0	45.5
Slovakia	54.5	52.4	50.8	44.5
Chile	53.9	50.2	49.1	44.1
Mexico	49.4	45.3	43.0	38.8
OECD average	**60.7**	**57.0**	**55.6**	**52.4**
No Response	**4.4**	**5.9**	**7.1**	**7.9**
Not Reached	**0.1**	**0.2**	**1.1**	**7.1**

Figure 4.9: Performance at the 2012 reading domain by cluster position.

Percentage Correct Responses at the PISA 2015 Reading Domain
By Cluster Position

	1	2	3	4
Finland	66.1	65.3	66.5	64.2
South Korea	66.2	64.8	63.5	62.5
Ireland	64.9	63.0	64.5	62.6
Estonia	66.1	63.9	63.8	61.1
Netherlands*	67.1	64.4	62.9	60.2
Japan	65.6	63.0	64.7	60.8
Germany	65.4	62.5	63.6	61.6
Norway	67.2	60.6	61.3	58.2
Belgium	64.8	63.0	60.3	57.9
Canada	64.7	61.3	60.9	57.8
Poland	63.0	60.3	62.0	58.7
New Zealand	64.7	59.2	61.0	56.7
Sweden	65.7	60.8	59.8	55.0
France	64.6	60.5	60.0	56.1
Spain	63.8	60.4	59.2	56.0
Czechia	63.8	59.6	59.5	56.5
United Kingdom	62.0	59.0	59.2	55.8
United States	62.6	58.9	58.3	55.7
Denmark	61.8	57.9	59.6	55.7
Italy	64.0	57.7	58.8	54.3
Latvia	61.8	59.2	58.4	54.6
Chile	62.0	58.2	58.4	55.2
Slovenia	61.6	59.1	57.4	54.0
Austria	61.9	57.6	57.5	54.9
Switzerland	62.4	58.1	57.7	53.2
Portugal	62.4	58.0	56.7	54.2
Australia	60.9	57.4	57.1	53.7
Hungary	60.0	56.7	56.6	53.4
Israel	60.5	55.3	57.1	53.2
Luxembourg	60.6	55.9	56.9	51.8
Iceland	59.2	55.7	55.3	53.0
Greece	60.1	52.2	56.1	51.9
Slovakia	56.4	51.5	53.3	49.8
Mexico	52.6	48.1	49.6	46.5
Turkey	49.6	44.7	45.6	42.0
OECD average	**62.6**	**58.8**	**59.0**	**55.7**
No Response	**3.2**	**4.4**	**4.5**	**5.7**
Not Reached	**0.3**	**1.2**	**0.4**	**0.9**

Figure 4.10: Performance at the 2015 reading domain by cluster position.

RMSD Percentage Correct Responses at the PISA Reading Domain
Across Cluster Positions

	2012	2015
Japan	3.1	1.8
South Korea	2.1	1.4
Ireland	1.8	1.0
Estonia	2.8	1.8
Finland	1.9	0.9
Netherlands*	3.3	2.5
Belgium	2.2	2.6
Germany	2.7	1.4
Poland	3.9	1.6
Canada	2.6	2.4
New Zealand	3.3	2.9
Norway	2.8	3.3
France	2.8	3.0
Czechia	2.4	2.6
United States	2.7	2.5
Latvia	Not OECD	2.6
United Kingdom	2.3	2.2
Sweden	2.9	3.8
Spain	3.4	2.8
Italy	3.8	3.5
Australia	2.9	2.5
Switzerland	2.9	3.3
Austria	1.9	2.5
Denmark	2.2	2.2
Israel	4.3	2.7
Hungary	2.8	2.3
Portugal	3.9	3.0
Luxembourg	3.2	3.1
Iceland	3.5	2.2
Slovenia	3.8	2.8
Greece	5.8	3.3
Chile	3.5	2.4
Slovakia	3.7	2.4
Turkey	3.5	2.7
Mexico	3.8	2.2
OECD average	**3.0**	**2.4**
No Response	**1.3**	**0.9**
Not Reached	**2.9**	**0.4**

Figure 4.11: RMSD at the reading domain across cluster positions.

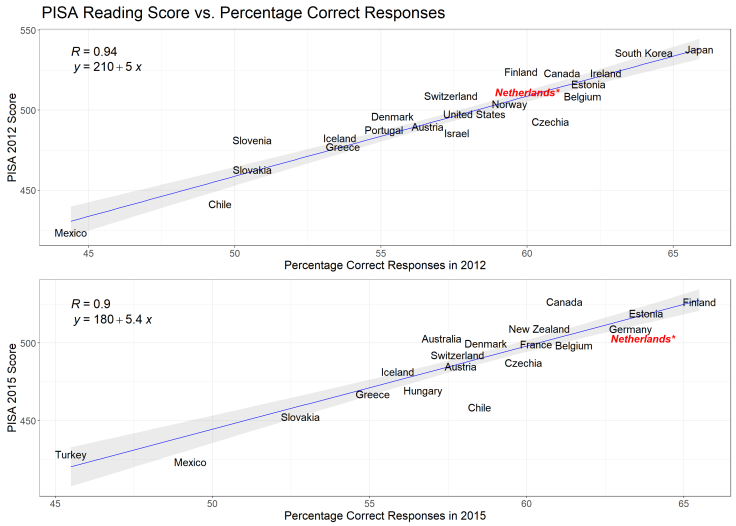

Figure 4.12: PISA reading score vs. percentage correct responses.

Figure 4.13, 4.14 and 4.15 illustrate developments in the percentage correct responses to reading items, the item-level percentage of non-response, and the percentage *not-reached* over the course of the PISA assessments for students in the Netherlands. The corresponding numbers can be found in the Appendix (Table A.4, A.5 and A.6).

The reading performance at the 2012 and 2015 cycles of both gender groups showed a similar rate of decline over the course of the assessment, with the female students consistently performing better at all cluster positions. Most noticeable, in 2012 the female students saw a steep drop in performance between the first and second cluster position (1: 67.8%; 2: 62.7%; 3: 61.9%; 4: 59.9%) while the male students' performance degraded more evenly (1: 63.0%; 2: 60.9%; 3: 57.3%; 4: 53.6%).

For both genders, the level of non-response in 2012 and 2015, did not noticeably change between the first and second cluster position, and saw a raise at the subsequent positions. In 2012, male students showed a higher percentage non-response at all cluster positions, and a steeper incline between the third and fourth position than the female students. At the 2015 cycle the average non-response and the trend over the course of the assessment was evenly matched between genders.

The level *not-reached* at the 2012 assessment developed equally for male and female students across cluster positions. The percentage *not-reached* at the first three cluster positions was close to zero for both genders. At the last position, male (1.1%) and female (1.3%) students saw a comparable incline in the proportion responses coded as *not-reached*. In 2015 the level *not-reached* was close to zero for female students at all four cluster positions and slightly elevated for male students at position 1, 2 and 3.

In 2012, all educational program groups saw a steady decline in reading performance over the course of the assessment. The drop in the percentage correct responses was least pronounced for the VWO group and most noticeable for the VMBO groups. Overall, the trends were comparable between PISA cycles. An exception formed the 2015 VWO group, which did not see a drop in performance when the first reading cluster was presented at the third position in a booklet. Furthermore, the VMBO BB group saw a slight incline in the percentage correct responses at the second cluster position of the 2012 assessment

(1: 36%; 2: 36.4%) and a slight decline in 2015 (1: 39.4%; 2: 38.9%). However, the difference between cycles was of small magnitude and could not be clearly distinguished from random error variation.

In general, low-performing educational groups showed a stronger increase and a higher overall level of non-response at both PISA cycles. The most notable differences between cycles were observed for the VMBO groups. For the 2012 VMBO GT students, a steep incline at the third cluster position followed by no change at the last position was observed. At the 2015 cycle, the VMBO GT group saw a steady increase in non-response that became steeper over the course of the assessment. At the 2012 assessment, the trend in non-response of VMBO KB students was monotonically positive across cluster positions, in 2015 the trend saw a decline at the second position, followed by a steep incline at the third and fourth position. At both cycles, the level of non-response observed for VMBO BB students declined noticeably at the second cluster position. However, in 2012 the non-response slightly increased at the third and strongly increased at the last position, while in 2015 it strongly increased at the third and declined at the last position.

In the 2012 assessment data, the level of *not-reached* saw an incline at the last cluster position across educational programs. The most noticeable increase was observed for the VMBO KB group (3: .0%; 4: 2.6%). The VMBO BB and VMBO GT groups moreover saw a slight incline in the percentage *not-reached* at the third cluster position. In 2015, with the exception of the VMBO BB group, the level of *not-reached* was close to zero for all groups at all cluster positions. For the VMBO BB students a comparable proportion of responses was coded as *not-reached* at the first, second and fourth cluster position, and a noticeably smaller proportion at the third position (1: 2.3%; 2: 2.4%; 3: 0.9%; 4: 2.2%).

At the third cluster position of the 2012 assessment, the performance of students in the third ESCS quantile (54.8%) dropped below the performance of students in the first (55.3%) and second quantile (58.9%), which saw an elevated performance at the third position. In 2015, the trend in performance did not noticeably differ between the ESCS groups, with higher ESCS students performing systematically better across cluster positions.

The level of non-response was on average higher at lower ESCS quantiles, however no systematic relationship between ESCS and the non-response pattern across cluster positions was observed. The most remarkable pattern was observed for the first ESCS quantile, which saw a steep incline in non-response at the last cluster position of the 2012 assessment and a noticeable drop in non-response at the second position in 2015. In 2012, the ESCS groups showed a similar pattern of the percentage *not-reached* across cluster positions. At the first three cluster positions, the level of *not-reached* was close to zero for all groups and noticeably elevated at the last position. In 2015, with the exception of an elevated percentage *not-reached* at the second and fourth cluster position of the third ESCS quantile, the ESCS groups showed a similar level of *not-reached* close to zero across all cluster positions.

On average, students native to the Netherlands showed a higher reading performance than first- and second-generation students at the 2012 and 2015 assessments. However, in 2012 their performance at the first cluster position was matched by the first-generation students. The first-generation group subsequently saw a strong decline in performance at the second cluster position and a slight decrease between the third and fourth position. In contrast, for the second-generation students an elevated performance at the second cluster

position, followed by a drop at the third and no change at the last position was observed. In 2015 the performance declined over the course of the assessment for all immigration groups. For the second-generation group the decline at the last cluster position was less steep compared to the native and first-generation groups.

The level of non-response of native students did not change between the first two cluster positions of the 2012 assessment and saw a gradual increase at the subsequent positions. Second-generation students saw an equal level of non-response at the first two cluster positions and a steep increase at the third, followed by a decline at the last position. The non-response in the first-generation group strongly increased between the first and second cluster position and between the third and fourth position. In 2015 the percentage non-response in the native and second-generation groups did not noticeably change between the first and second cluster position and saw an incline at the third and fourth position. The incline was steeper for the second-generation than the native students. The level of non-response of the first-generation group varied strongly across cluster positions, but the variation could not be clearly distinguished from random error due to the small group size.

In the 2012 assessment data, the level of *not-reached* was elevated from zero at the last cluster position for all immigration groups. Second-generation students moreover saw an increased level at the second position. In 2015 the proportion responses coded as *not-reached* was elevated at the second cluster position for the native and the first-generation students. The second-generation group saw a decreased level of *not-reached* at the second cluster position, followed by a gradual increase at the subsequent positions.

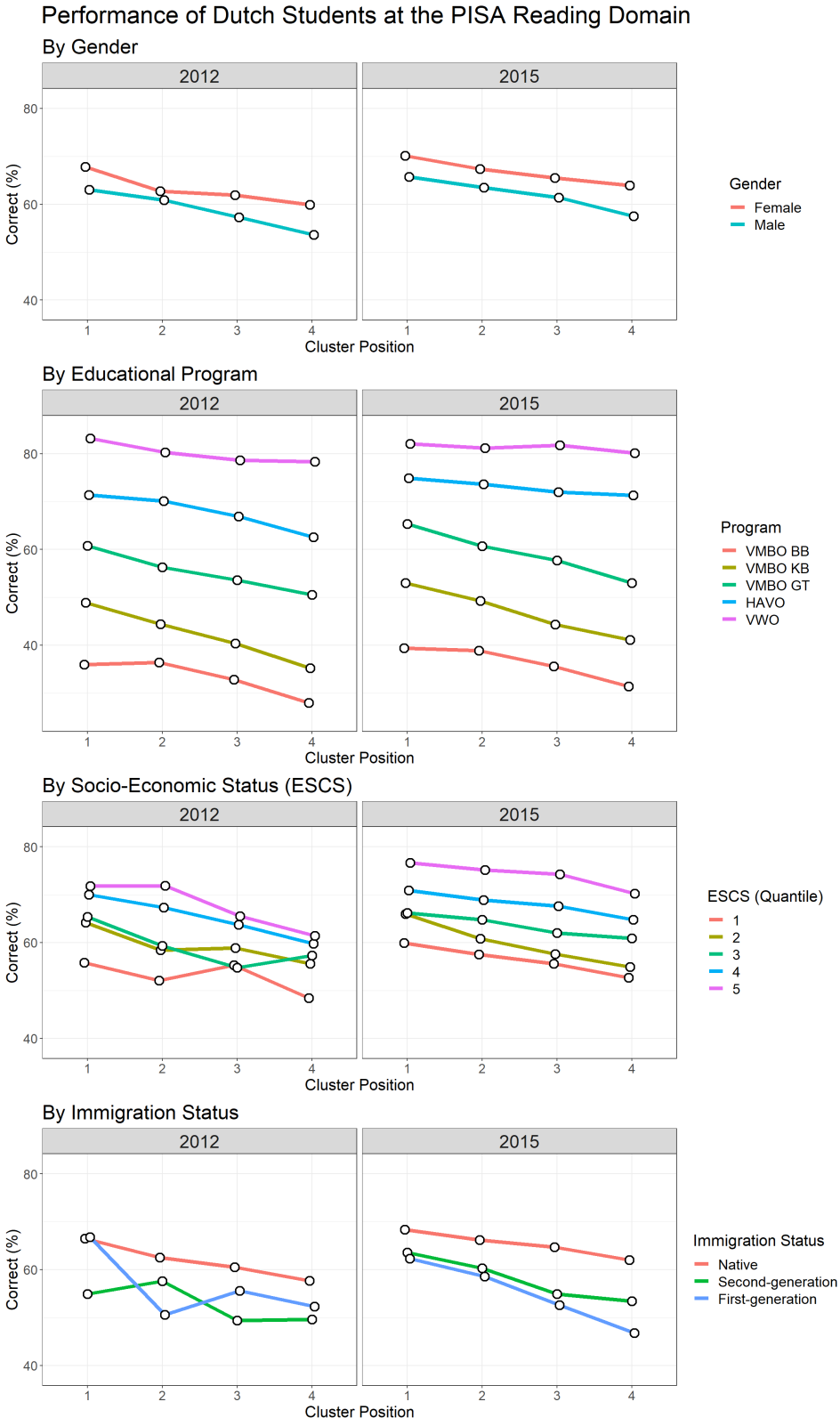

Figure 4.13: Performance of Dutch students at the PISA reading domain.

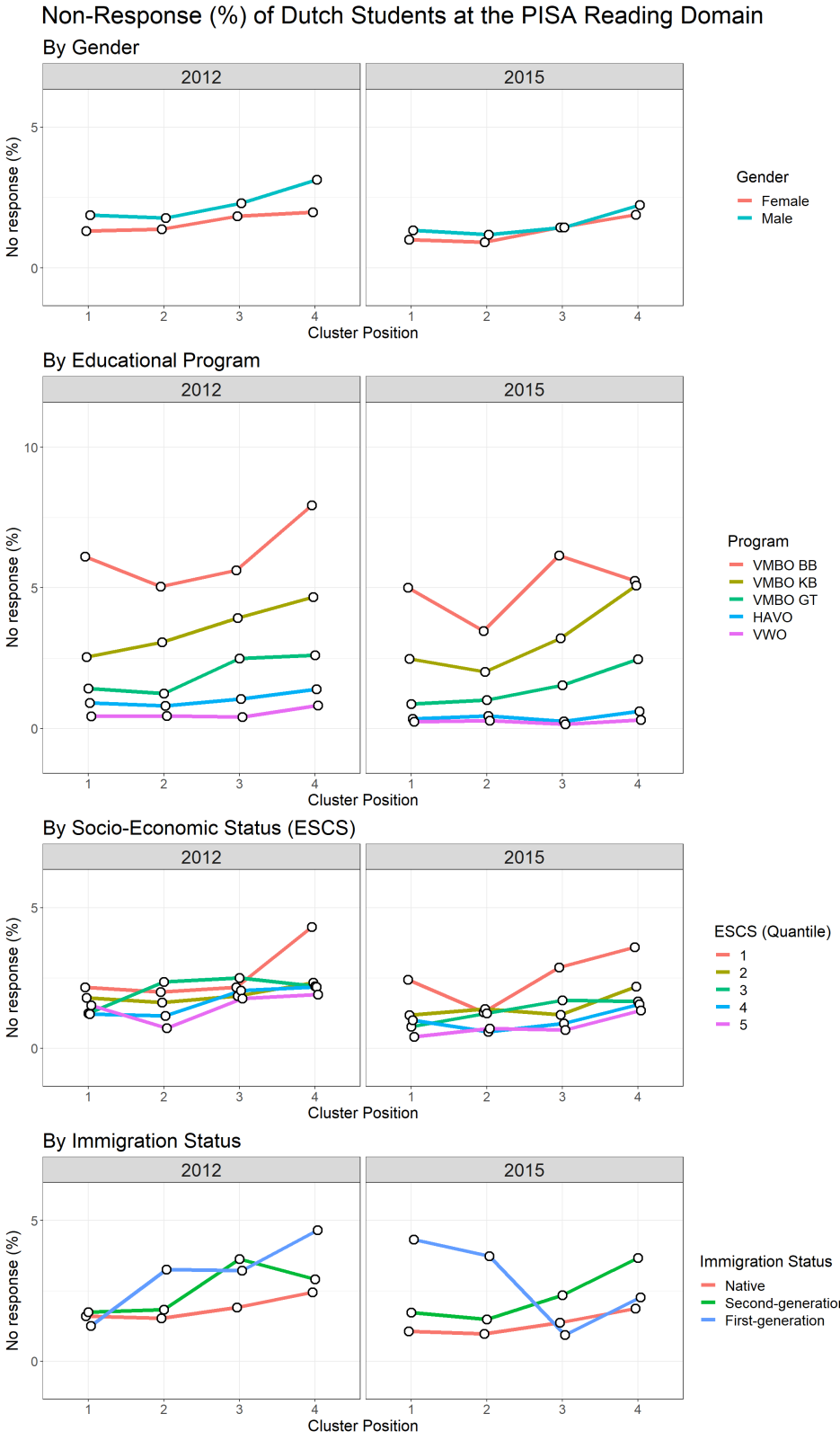

Figure 4.14: Non-Response (%) of Dutch students at the PISA reading domain.

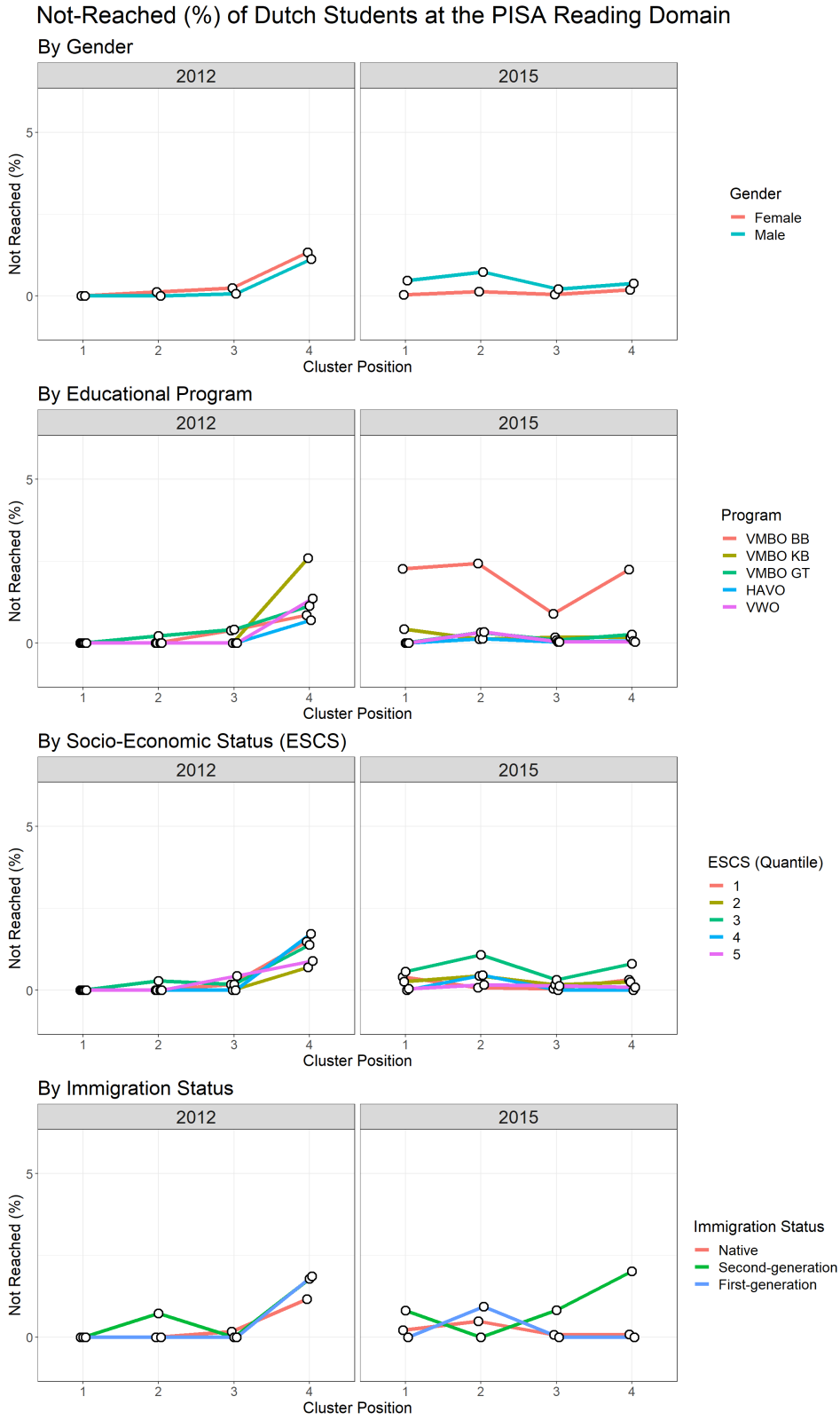

Figure 4.15: Not-Reached (%) of Dutch students at the PISA reading domain.

4.3 Science

In Figure 4.16, 4.17 and 4.18 the percentage correct responses to science items of students in OECD countries is shown for the 2012, 2015 and 2018 PISA cycles. The numbers are computed per cluster position to provide insight into the development of the students' test performance over the course of the PISA assessment. The figures furthermore contain the OECD-average percentages for item-level non-response and for response data coded as *not-reached*.

The OECD-average science performance gradually decreased over the course of the 2012 PISA assessment. With the exception of Finland (1: 60.6%; 2: 60.9%; 3: 58.1%; 4: 57.5%) and the Netherlands (1: 59.2%; 2: 59.8%; 3: 57.1%; 4: 53.8%) which saw a slightly elevated performance at the second cluster position, the negative trend in performance was consistent across countries. In 2015 a different trend was observed: the country-level performance consistently dropped at the second science cluster in a booklet (cluster position 2 or 4) compared to the respective first cluster (cluster position 1 or 3). Moreover, the performance was on average lower when the science clusters were presented at the end of the assessment (3: 49.8%; 4: 46.8%) compared to booklets in which the science clusters were placed at the beginning (1: 51%; 2: 48%).

As in 2015, the country-level performance dropped noticeably at the second science cluster (cluster position 2 and 4) at the 2018 assessment. The decline in performance was especially pronounced between the first and second cluster position (1: 52.4%; 2: 42.5%), which can be translated to an average drop of $(52.4 - 42.5) \cdot 4.5 = 44.6$ points in the PISA science score at the second cluster position of a booklet (the regression slope in Figure 4.20 serves as the translation factor). At the previous cycles, the decline between the first and the second cluster position translated to a drop of 6.9 points (2012) and 15.3 points (2015) on the PISA scale. The decline between the third and fourth cluster position in 2018 was less severe (3: 49.3%; 4: 44%; 23.9 points) but noticeably stronger than in the previous cycles (2012: 13.3 points; 2015: 15.3 points). The pattern was consistently observed across countries and more strongly pronounced at the low-performing OECD spectrum.

As shown in Figure 4.19, South Korea (1.3%) and the Netherlands (2.8%) saw the least variation in performance between cluster positions of the top 10 science OECD countries at the 2018 assessment. Strong variations were observed at the bottom end of the performance spectrum (Colombia: 5.9%; Mexico: 5.5%; Turkey: 4.8%; Greece: 5.1%) and for countries that performed around the OECD median (Sweden: 7.3%; France: 5.3%; Norway: 5.2%). Of the top 10 group, the positioning of science clusters in the 2018 assessment had the strongest effect on the observed performance in Germany (4%) and Japan (3.9%). The average variation in performance between cluster positions was lower in 2012 (2.2%) and 2015 (1.6%) compared to 2018 (4%). For the 2012 and 2015 assessments, no clear relationship between the countries' performance level and the country-average effect of cluster positioning was observed.

The relationship between performance and PISA score illustrated in Figure 4.20 indicates that for 2018, the Netherlands scored noticeably lower than expected on the official scale given the average percentage correct responses in the country. The derivation can be explained by the high (relative to previous cycles) proportion of (low-performing) Dutch students that were presented the UH test form. Across OECD countries the correlation

between the percentage correct responses and the PISA score was very high at all three cycles (2012: .95; 2015: .96; 2018: .95).

At the 2012 PISA cycle, the OECD-average percentages of item-level non-response gradually increased over the course of the assessment. The proportion responses coded as *not-reached* was close to zero at the first two cluster positions, followed by a slight incline at the third, and a steep incline at the last position of the assessment. In 2015 and 2018, the percentages item-level non-response and *not-reached* were elevated in the second science cluster in a booklet (cluster position 2 and 4). The shifts in non-response across cluster positions were of a similar magnitude at the three investigated cycles, however the proportion responses coded as *not-reached* saw a very strong increase at the second and fourth cluster position of the 2018 assessment (1: .9%; 2: 12.5%; 3: 1.2%; 4: 5.8%). Note that in 2015 and 2018, responses coded as *not-reached* were scored as incorrect, whereby the level of *not-reached* is likely to correlate with the observed percentage correct responses. The following analysis provides further insight into patterns of missingness within the Netherlands at the PISA science domain.

Percentage Correct Responses at the PISA 2012 Science Domain
By Cluster Position

	1	2	3	4
Japan	65.7	63.4	61.6	59.6
Estonia	63.5	61.8	61.6	58.0
South Korea	62.8	62.2	60.0	58.3
Czechia	62.0	61.2	58.9	57.1
Germany	62.2	59.8	59.1	56.8
Finland	60.6	60.9	58.1	57.5
Poland	62.2	59.8	58.2	53.9
Netherlands*	59.2	59.8	57.1	53.8
Ireland	58.4	58.1	57.5	54.9
New Zealand	59.1	57.1	57.7	53.2
Canada	58.7	58.1	56.0	53.5
Belgium	59.3	57.1	55.4	54.3
Australia	58.3	56.8	55.4	52.7
Austria	57.6	56.7	54.9	53.3
United Kingdom	57.5	56.8	53.8	52.3
France	58.0	55.4	52.8	52.1
Spain	56.8	55.6	54.0	50.3
Switzerland	55.7	55.4	53.6	51.1
Hungary	56.6	55.0	53.1	50.6
Italy	57.1	55.5	52.5	48.9
Norway	57.1	53.1	52.8	49.4
Slovenia	56.4	54.6	51.6	48.5
United States	54.4	54.1	51.7	49.9
Sweden	55.3	53.9	50.6	48.2
Luxembourg	55.3	51.7	50.4	48.9
Portugal	54.3	52.5	50.4	46.0
Denmark	52.3	51.1	49.8	48.3
Iceland	52.3	52.2	48.5	45.3
Slovakia	52.7	50.0	47.2	45.9
Israel	53.0	50.5	47.7	44.6
Greece	52.1	50.3	45.6	40.2
Chile	50.2	48.0	45.3	41.6
Turkey	47.9	46.9	44.7	41.6
Mexico	40.9	39.5	37.0	33.0
OECD average	**55.3**	**54.0**	**51.9**	**49.4**
No Response	**3.8**	**5.0**	**6.3**	**7.5**
Not Reached	**0.0**	**0.2**	**1.2**	**6.8**

Figure 4.16: Performance at the 2012 science domain by cluster position.

Percentage Correct Responses at the PISA 2015 Science Domain
By Cluster Position

	1	2	3	4
Japan	58.5	56.7	59.0	55.9
Estonia	57.9	55.8	58.4	56.4
Finland	57.3	56.0	57.1	55.5
Netherlands*	55.0	53.3	53.4	50.9
Germany	54.2	52.3	53.6	51.4
Canada	54.5	51.9	53.2	51.2
South Korea	52.7	52.3	53.0	51.5
New Zealand	54.8	51.4	52.9	48.3
Belgium	52.7	50.3	51.9	49.5
Czechia	51.8	49.7	52.3	49.5
United Kingdom	53.1	50.6	51.2	48.3
Poland	51.4	49.8	51.8	48.9
Ireland	50.9	50.0	51.0	49.1
Austria	51.7	49.1	50.7	48.4
Slovenia	51.9	49.4	50.1	48.0
Switzerland	51.9	48.7	51.0	47.1
Spain	51.8	48.2	50.9	47.5
United States	51.3	49.1	50.0	47.7
Australia	52.0	49.0	50.4	46.5
Norway	52.5	48.1	50.1	46.3
France	52.1	47.1	50.3	46.5
Sweden	50.4	45.9	50.1	46.9
Italy	51.6	46.9	49.6	44.8
Latvia	48.3	46.9	49.7	46.3
Denmark	49.4	46.1	48.9	46.8
Hungary	48.9	46.9	47.9	45.2
Portugal	49.9	45.3	48.0	44.2
Luxembourg	48.6	44.9	46.8	44.6
Iceland	46.7	44.7	44.1	41.5
Israel	46.9	42.1	45.0	40.5
Slovakia	47.6	42.6	43.1	39.1
Chile	45.1	41.1	43.6	40.5
Greece	45.2	39.0	44.1	40.2
Turkey	36.6	33.2	35.7	32.0
Mexico	36.2	32.3	34.3	30.9
OECD average	**51.0**	**48.0**	**49.8**	**46.8**
No Response	**2.3**	**3.3**	**2.9**	**3.9**
Not Reached	**0.9**	**1.8**	**1.0**	**1.7**

Figure 4.17: Performance at the 2015 science domain by cluster position.

Percentage Correct Responses at the PISA 2018 Science Domain
By Cluster Position

	1	2	3	4
Estonia	59.9	51.2	59.0	53.9
Netherlands*	59.1	53.4	55.1	51.4
South Korea	55.7	53.3	54.1	52.1
Japan	58.5	49.2	56.7	50.7
Finland	57.3	49.5	56.3	52.0
Czechia	56.8	48.9	53.6	49.1
Poland	56.7	48.3	53.6	49.5
Canada	56.6	47.7	52.5	48.1
United Kingdom	54.4	48.6	50.6	46.6
New Zealand	54.5	45.9	51.5	47.7
Germany	54.9	44.5	52.1	48.0
Australia	54.7	47.4	50.8	46.3
Slovenia	53.6	47.5	49.7	46.6
Ireland	53.7	46.8	49.8	46.5
Austria	53.2	45.9	50.0	46.9
Belgium	55.3	43.6	51.4	45.3
United States	53.2	43.4	49.5	46.5
Spain	52.7	41.4	49.6	43.3
Switzerland	52.2	40.5	49.9	43.6
Latvia	52.0	42.2	48.2	43.5
Hungary	49.9	43.2	47.6	44.9
Portugal	53.0	39.9	49.3	43.1
Lithuania	49.8	43.1	47.8	43.7
Norway	52.6	40.4	48.0	40.4
Denmark	50.5	40.2	47.8	42.5
Sweden	54.2	36.7	50.4	39.5
Italy	50.4	40.2	47.8	41.1
France	50.8	37.6	48.0	41.1
Slovakia	48.6	39.9	45.5	40.8
Luxembourg	48.6	38.2	45.5	38.4
Iceland	48.8	36.2	45.2	38.3
Turkey	47.7	35.5	45.5	39.6
Israel	47.6	36.7	44.1	38.2
Chile	45.6	31.9	42.3	35.0
Greece	43.6	31.1	42.7	35.9
Mexico	38.1	24.4	36.2	29.4
Colombia	38.3	23.7	35.4	27.3
OECD average	**52.4**	**42.5**	**49.3**	**44.0**
No Response	**2.2**	**3.6**	**3.4**	**4.4**
Not Reached	**0.9**	**12.5**	**1.2**	**5.8**

Figure 4.18: Performance at the 2018 science domain by cluster position.

RMSD Percentage Correct Responses at the PISA Science Domain
Across Cluster Positions

	2012	2015	2018
Estonia	2.0	1.1	3.6
Japan	2.2	1.3	3.9
Finland	1.5	0.7	3.2
South Korea	1.8	0.6	1.3
Netherlands*	2.3	1.5	2.8
Czechia	1.9	1.2	3.3
Germany	1.9	1.1	4.0
Poland	3.0	1.2	3.3
Canada	2.0	1.3	3.6
New Zealand	2.2	2.4	3.3
Ireland	1.4	0.8	2.9
Belgium	1.9	1.3	4.7
United Kingdom	2.1	1.7	2.9
Australia	2.1	2.0	3.3
Austria	1.7	1.3	2.9
Slovenia	3.0	1.4	2.7
Spain	2.4	1.8	4.6
United States	1.8	1.3	3.6
Switzerland	1.8	1.9	4.7
France	2.3	2.3	5.3
Norway	2.7	2.3	5.2
Hungary	2.2	1.4	2.6
Italy	3.1	2.6	4.3
Sweden	2.8	2.0	7.3
Portugal	3.1	2.2	5.1
Denmark	1.5	1.4	4.1
Latvia	Not OECD	1.3	3.9
Luxembourg	2.4	1.6	4.5
Lithuania	Not OECD	Not OECD	2.8
Iceland	2.9	1.9	5.1
Slovakia	2.6	3.0	3.5
Israel	3.1	2.5	4.4
Chile	3.2	1.9	5.5
Greece	4.6	2.6	5.1
Turkey	2.4	1.9	4.8
Mexico	3.0	2.0	5.5
Colombia	Not OECD	Not OECD	5.9
OECD average	**2.2**	**1.6**	**4.0**
No Response	**1.4**	**0.6**	**0.8**
Not Reached	**2.8**	**0.4**	**4.7**

Figure 4.19: RMSD at the science domain across cluster positions.

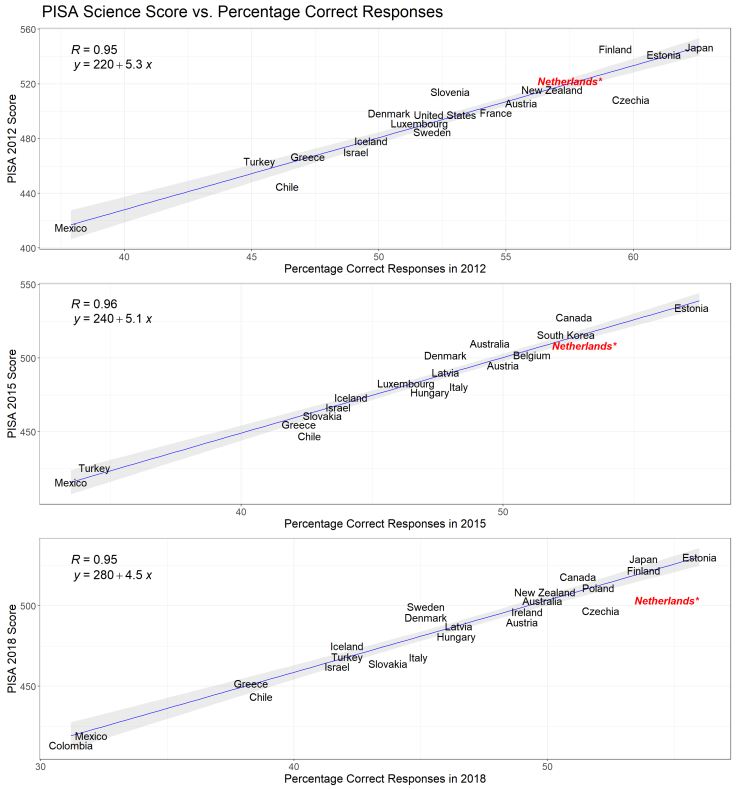

Figure 4.20: PISA science score vs. percentage correct responses.

Figure 4.21, 4.22 and 4.23 illustrate developments in the percentage correct responses to science items, the item-level percentage of non-response, and the percentage *not-reached* over the course of the PISA assessments for students in the Netherlands. The corresponding numbers can be found in the Appendix (Table A.7, A.8 and A.9).

In 2015, female and male students in the Netherlands saw a slight increase in the percentage correct responses at the second cluster position and a decrease in performance at the subsequent positions. Most noticeably, female students saw a strong decline in performance between position 3 (58.7%) and 4 (53.2%) while for male students the most severe incline was observed between position 2 (61.6%) and 3 (57.4%). At the 2015 assessment, both gender groups saw a decline in performance at the second science cluster in a booklet (position 2 and 4) and no change between position 2 and 3. In 2018, the trend in, and the average of the percentage correct responses did not noticeably differ between genders over the course of the assessment. Both groups saw a decline in performance at the second science cluster in a booklet (position 2 and 4) and a lower performance when the science clusters were presented at the end of the assessment (position 3 and 4) compared to clusters placed at the beginning (position 1 and 2).

For female students a gradual incline in the percentage non-response was observed over the course of the 2012 assessment, with the steepest increase observed at the fourth cluster position. The level of non-response of the male group did not differ for clusters in the same booklet (position 1 and 2, or 3 and 4), but was noticeably higher for cluster presented at the end of the assessment. In 2015, the level of non-response did not noticeably differ between genders. Both groups saw a constant level of non-response at the first three cluster positions and slightly elevated non-response at the last position. At the 2018 cycle, a monotonically increasing level of non-response over the course of the assessment was observed for the female group. The male students saw elevated non-response at cluster position 2 and 4.

For both genders, a constant percentage *not-reached* was observed at the first three cluster positions of the 2012 assessment. At the fourth position, the percentage was slightly higher for the female group (female: 1%; male: .6%). In 2015, male and female students saw a slightly elevated level of *not-reached* at the second and fourth cluster position. In the 2018 assessment data, a strongly elevated proportion responses coded as *not-reached* was observed for the female group at the second and fourth cluster position (1: .1%; 2: 3.8%; 3: .8%; 4: 3.1%). For male students a similar pattern was observed, however the male group showed a less steep incline at the last cluster position (1: .5%; 2: 3.6%; 3: .5%; 4: 1.5%).

In 2012, the percentage correct responses of the VWO group increased between the first and third cluster position and declined noticeably at the last position. For HAVO students the performance was noticeably higher at science clusters that were placed at the beginning of a booklet (cluster position 1 and 2). The performance of the HAVO group moreover increased between the first and second cluster position and slightly declined between the third and fourth position. The performance of the VMBO groups monotonically declined over the course of the assessment. At the 2015 assessment, the VWO and VMBO BB groups did not see noticeable drops in performance across cluster positions. The performance of the HAVO and VMBO GT groups degraded over the course of the assessment, for the VMBO KB group an elevated percentage correct responses at the first and third cluster

position was observed. In 2018, all educational groups saw a steep drop in performance at the second and fourth cluster position. With the exception of the VMBO BB group, an overall higher performance in booklets that placed the science clusters at the beginning of the assessment (cluster position 1 and 2) was observed. Note that the results for the 2018 VMBO BB group are subject to strong random variation due to the small sample size.

The level of non-response differed noticeably between educational programs. In general, low-performing groups showed a stronger incline and a higher overall level of non-response at all three PISA cycles. The most notable difference between cycles was observed for the VMBO KB and VMBO GT groups. In 2012 and 2018, the level of non-response in the VMBO GT group increased gradually over the course of the assessment, in 2015 no change between cluster positions was observed. The VMBO KB group saw a steep incline in the percentage non-response after the second cluster position of the 2012 assessment. In 2015 the non-response of the VMBO KB students declined slightly between cluster position 1 and 3, and was slightly increased at the last position. At the 2018 cycle, the development in non-response over the course of the assessment was similar for the VMBO BB and VMBO KB groups, with noticeably elevated levels of non-response at the second and fourth cluster position.

Across educational programs, the level of *not-reached* was close to zero at the first three cluster positions of the 2012 assessment. At the last cluster positions all programs but the VMBO BB group saw a slight incline in the proportion responses coded as *not-reached*. In 2015, the HAVO, VMBO GT and VMBO BB groups saw an elevated percentage *not-reached* at the second and fourth cluster position. For the VWO group a constant level of *not-reached* was observed over the course of the 2015 assessment. The VMBO KB group saw an increase in the level of *not-reached* at the second cluster position. Moreover, for the VMBO KB students the proportion responses coded as *not-reached* was on average lower at science clusters that were placed at the end of the assessment (cluster position 3 and 4).

In 2018, with the exception of the VMBO KB group, the percentage *not-reached* of students in the Netherlands was noticeably elevated at the second and fourth cluster position of the assessment. The pattern was less pronounced in the VMBO GT group and strongly pronounced in the VMBO BB, HAVO and VWO groups. For the VMBO KB group, a monotone incline in the level of *not-reached* over the course of the assessment was observed.

Students with a higher ESCS systematically performed better across cluster positions at all three PISA cycles. No systematic relationship between the students' ESCS and their performance pattern across cluster positions was observed. Most noticeable, the first quantile saw a strong decline at the last cluster position in 2012 (1: 52.9%; 2: 51.5%; 3: 49.4%; 4: 43.9%) and 2018 (1: 51.8%; 2: 47.1%; 3: 47%; 4: 40.3%). The level of non-response was on average higher at lower ESCS quantiles, however no systematic three-way interaction between the level of non-response, the students' ESCS and the cluster position was found. Similarly, the ESCS did not systematically affected the pattern *not-reached* across cluster positions.

Finally, students native to the Netherlands showed a higher science performance than first- and second-generation students across all cluster positions at all three PISA cycles. The percentage correct responses of native students saw a steady decline after the second cluster position of the 2012 assessment. In 2015 and 2018, a drop in performance of

native students at the second and fourth cluster position was observed. The pattern was more strongly pronounced in 2018. Moreover, the average performance was higher when science clusters were presented at the beginning (cluster position 1 and 2) compared to a placement at the end (cluster position 3 and 4) of the assessment.

At the 2012 and 2015 assessments, the performance of second-generation students declined monotonically across cluster positions. For the first-generation group an elevated performance at the second cluster position was observed. In 2018, the performance of the first- and second-generation groups was noticeably degraded at the second science cluster in a booklet (cluster position 2 and 4). Due to the small sample size of the 2015 and 2018 first-generation groups, the results must be interpreted with caution.

The level of non-response in the native group was overall low across PISA cycles. In 2012, 2015 and 2018 it saw the strongest incline at the third, fourth and second cluster position, respectively. The second-generation group saw elevated levels of non-response at the second and fourth cluster position of the 2012 and 2018 assessments. The pattern was more strongly pronounced in 2018. In 2015, the non-response of the second-generation students declined steadily between the first and third cluster position and increased at the last position. The first-generation group saw a decline in non-response at the second cluster position of the 2012 assessment, followed by a steep incline at the subsequent positions. In 2015 and 2018, the percentage non-response of the first-generation students steeply increased across cluster positions. Overall, the second- and first generation groups showed a higher level of non-response than the native group across all PISA cycles.

In 2012, the level of *not-reached* was close to zero at the first three cluster positions for all immigration groups. At the fourth cluster position, the percentage *not-reached* increased noticeably stronger for the second-generation (3.1%) and first-generation (2%) students compared to the native group (.6%). The native students saw a slight elevation in the level of *not-reached* at the last position of the 2015 assessment. For the second-generation students, the percentage *not-reached* was noticeably higher at science clusters at the beginning of the 2015 assessment (1: 2.3%; 1: 3.0%) compared to clusters that were presented at the end of the assessment (3: .6%; 4: .9%). In 2015, the first-generation group saw a steep incline at cluster position 3 and 4. At the 2018 cycle, for all immigration groups an elevated level of *not-reached* at the second and fourth cluster position was observed. The increase at the second (native: 3.2%; second-generation: 6.9%; first-generation: 15.3%) and fourth (native: 1.9%; second-generation: 3.4%; first-generation: 14.8%) cluster position was the the most pronounced in the first-generation group, and stronger for second-generation than native students.

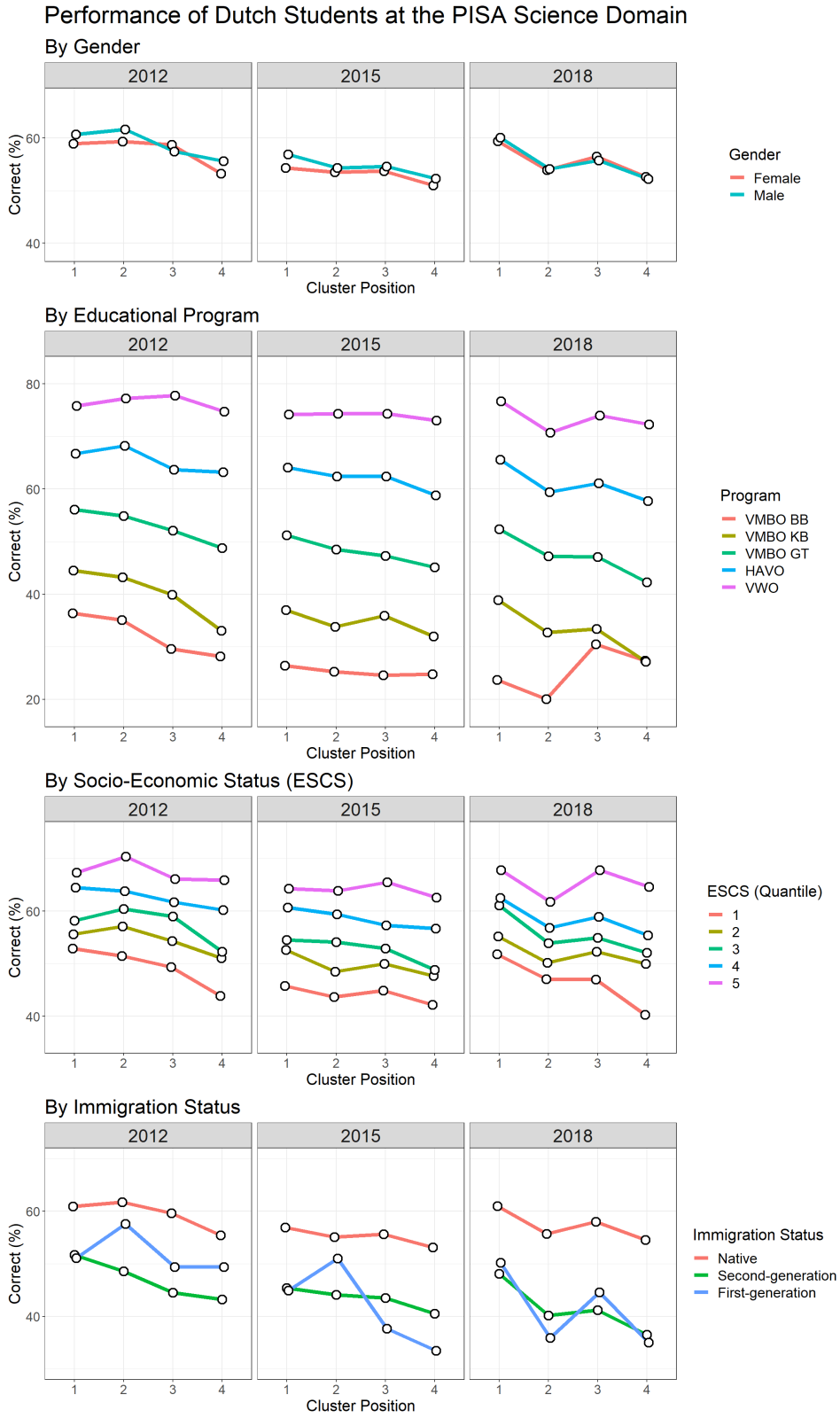

Figure 4.21: Performance of Dutch students at the PISA science domain.

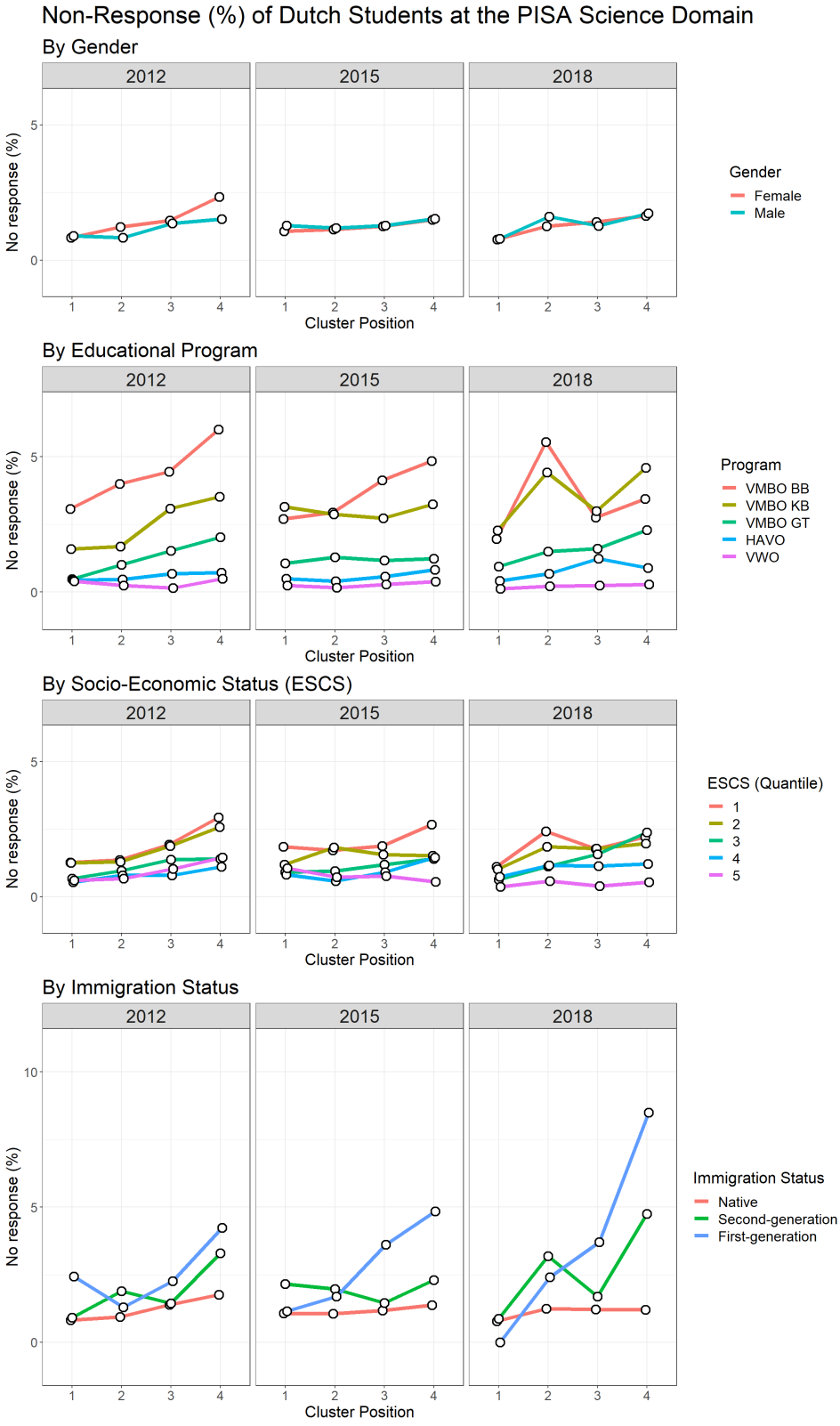

Figure 4.22: Non-Response (%) of Dutch students at the PISA science domain.

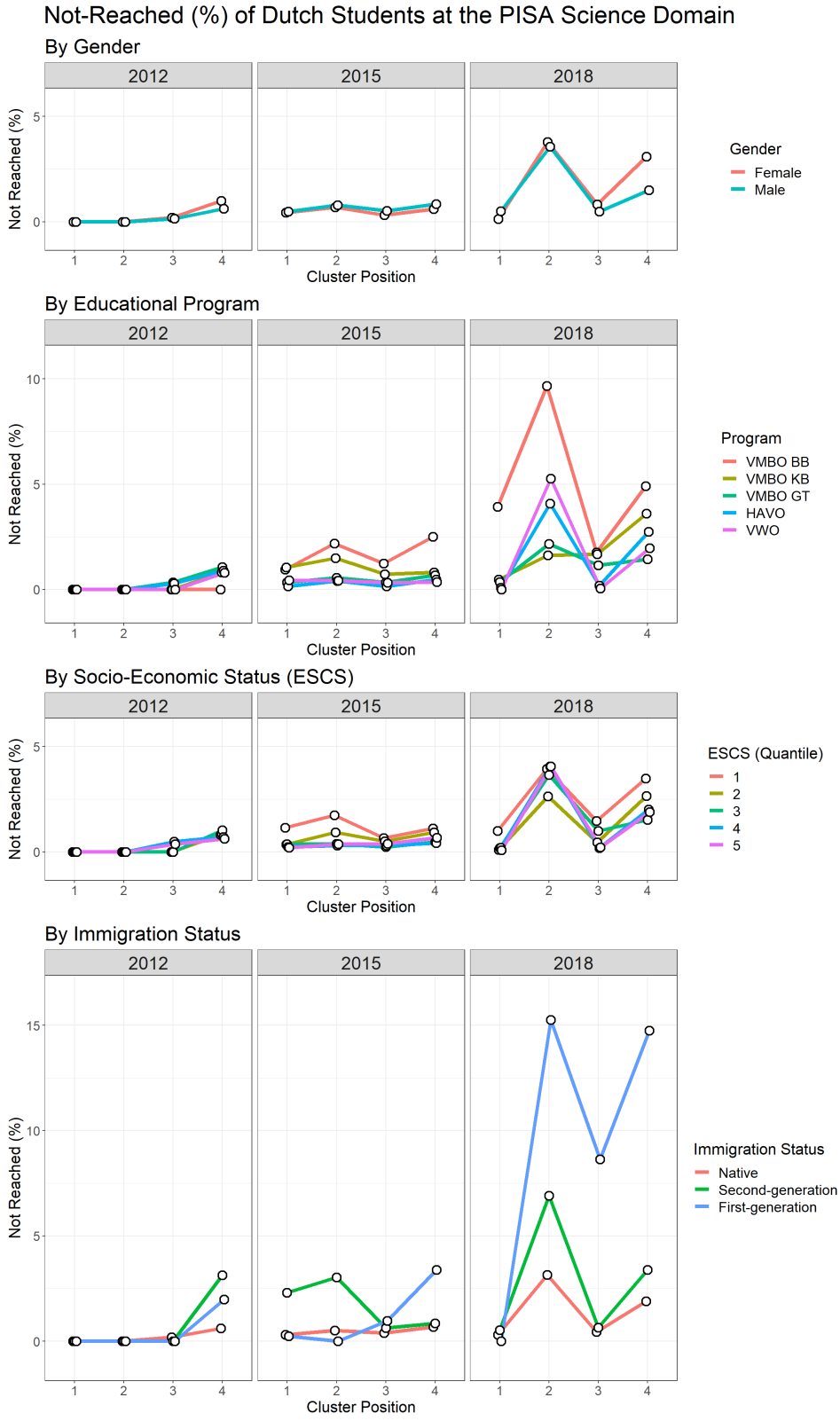

Figure 4.23: Not-Reached (%) of Dutch students at the PISA science domain.

5. Conclusion

The development of test performance over the course of the PISA assessment differed substantially between PISA cycles and across OECD countries. In 2012, a monotone decline in performance between subsequent cluster positions was observed for the mathematics, reading and science domains. The trend was consistent across countries and is in line with the gradual performance decline expected in progressing low-stake educational assessments (List et al., 2017; M. Wu, 2010).

At the 2015 and 2018 cycles, a different pattern in the performance trend was observed: students were on average less likely to provide a correct response to mathematics (2018), reading (2015) and science (2015 and 2018) items when the corresponding cluster was preceded by a cluster of the same domain in the booklet. Concretely, students performed worse after they were presented a 30-minutes cluster of items measuring the same domain, compared to two clusters of items covering a different domain. The pattern was most pronounced at the science domain and in low-performing countries. It did moreover consistently occur in combination with the first pattern, i.e., the average performance across clusters that measure the same domain was lower when the block of clusters was placed at the end of a booklet.

The results are indicative of cluster order effects that possibly interact with cognitive factors related to the students' invested effort during the test-taking process. For example, it is plausible that the students' level of boredom (Pekrun et al., 2010), task enjoyment (M. A. Lindner et al., 2019; Penk et al., 2014) or self-control (C. Lindner et al., 2018) shifts after answering a 30-minutes series of questions covering the same domain. Further research is required to explore the relationship between order effects, item characteristics and cognitive student-level factors.

Note that in contrast to the following PISA cycles, in the 2012 assessment design domain-specific clusters were not grouped to be presented consecutively to students. As a consequence, a mathematics cluster was not necessarily directly followed by a second

mathematics cluster in booklet. Furthermore, reading and science clusters were never directly followed or preceded by a cluster of the same type. The 2012 findings therefore support the hypothesis that the observed cluster order effects interact with the number of, or the continuous time spend on, items of the same domain. A further contributing factor for the discrepancy in the observed performance patterns between PISA cycles can possibly be found in the transition from paper- to computer-based presentation modes at the 2015 assessment.

The OECD-average variation in PISA score between cluster positions in a booklet declined between the 2012 cycle (mathematics: 8.3 points; reading: 15 points; science: 11.7 points) and the 2015 cycle (mathematics: 5.2 points; reading: 13 points; science: 8.2 points), and saw an incline at the 2018 cycle (mathematics: 5.8 points; science: 18 points). The trend was not consistent across countries. For example, the Netherlands saw a steady decline in the score variation at the mathematics and reading domains, i.e., the differences in test performance between cluster positions became less severe at later PISA cycles. In contrast, for Sweden a stronger variation in mathematics and reading score was observed at later PISA cycles. At the science domain, both countries showed a similar trend in score variation, however the magnitude of the changes in variation across PISA cycles differed greatly between Sweden (2012: 14.8 points; 2015: 10.2 points; 2018: 32.9 points) and the Netherlands (2012: 12.2 points; 2015: 7.7 points; 2018: 12.6 points).

The two-country example shows that comparisons between countries and/or cycles can be challenging, given that differences in the level of score variation can indicate that the extend to which the PISA score is affected by factors that are not part of the latent constructs measured by the PISA study (i.e., the students' proficiency levels) differs between the compared groups. PISA scores that are affected to a different extend by additional factors (e.g., fatigue, boredom or effort) cannot be placed on the same measurement scale, unless the influence of the additional factors is fully controlled for. At several domain-cycle combinations, low-performing countries tended to exhibit a stronger score variation than countries placed at the upper half of the corresponding domain-cycle performance ranking. The finding is in line with the results of Q. Wu et al., 2019, who found that the correlation between the students' performance level and their persistence during PISA assessments was noticeably stronger in low-performing countries. The country-level performance did however not comprehensively explain the differences in score variation that were observed across the investigated PISA cycles and countries.

The Netherlands-specific analyses revealed no substantial differences between male and female students in the development of test performance over the course of the PISA assessments. The most noticeable difference between gender groups was found in the proportion responses coded as non-response or *not-reached* at the end of the assessments, which tended to be higher for female students. The difference was most pronounced at the 2018 science domain. A possible explanation can be found in the higher level of motivation exhibited by female students at low-stake assessments (Butler & Adams, 2007; Eklöf, 2007; Kornhauser et al., 2014), which could be linked to an earlier onset of rapid-guessing behaviour of male students in the PISA assessments (M. A. Lindner et al., 2019). The hypothesis can be further explored by modelling the relationship between item-level response times and the trend in test performance for the gender groups.

Moderate overlap was found between educational programs and ESCS groups in the

performance development and for the patterns of missing response data over the course the PISA assessments. The results are indicative of a correlation between the students' followed educational program and their ESCS index. A plausible explanation can be found in a confounding role of the students' proficiency level, which is likely correlated with both the followed educational program and the ESCS. The differences between groups were more distinct in the program-specific analyses, which, following the aforedescribed hypothesis, would indicate that the students' educational program correlates stronger with the confounding variable than their ESCS index. The results moreover provided further support for the role of the students' proficiency level in predicting the development of their test performance over the course of the PISA assessments. However, similar to the OECD country-level analyses, the measured performance could only explain part of the variation in order effects between subpopulations of Dutch students.

Students native to the Netherlands showed a consistently higher performance and less score variation across cluster positions than first- and second-generation students. The finding was moreover reflected by the proportions responses coded as non-response or *not-reached* that remained more stable for the native students over the course of the PISA assessments. Due to the small size of the first- and second generation groups, the results should however be interpreted with caution.

Finally, it is possible that ignoring missing response data led to an underestimation of the performance decline during the 2012 assessment. The correlation between the obtained performance measure and the official PISA country scores was however very high across all domains, which indicates that the effect of missingness on the performance measure was similar across the investigated countries. Nevertheless, given the variety of ways in which missing response data in the PISA assessments can be handled (e.g., affecting item parameters, included as covariates in the estimation of plausible values, and/or utilized conditional on indices of motivation or rapid-guessing behavior), the impact of different treatments of missing data on order effects and performance measures can be a topic of future research.

Bibliography

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.

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.

Eklöf, H. (2007). Test-Taking Motivation and Mathematics Performance in TIMSS 2003 [Publisher: Routledge _eprint: https://doi.org/10.1080/15305050701438074]. *International Journal of Testing*, *7*(3), 311–326. https://doi.org/10.1080/15305050701438074

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

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

Kornhauser, Z. G. C., Minahan, J., Siedlecki, K. L., & Steedle, J. T. (2014). *A Strategy for Increasing Student Motivation on Low-Stakes Assessments* [Publication Title: Council for Aid to Education]. Council for Aid to Education. Retrieved December 15, 2021, from https://eric.ed.gov/?id=ED582123

Lindner, C., Nagy, G., & Retelsdorf, J. (2018). The need for self-control in achievement tests: Changes in students' state self-control capacity and effort investment [Place: Germany Publisher: Springer]. *Social Psychology of Education: An International Journal*, *21*(5), 1113–1131. https://doi.org/10.1007/s11218-018-9455-9

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The Onset of Rapid-Guessing Behavior Over the Course of Testing Time: A Matter of Motivation and Cognitive Resources. *Frontiers in Psychology*, *10*, 1533. https://doi.org/10.3389/fpsyg.2019.01533

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

List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., & Nagy, G. (2017). Performance decline in low-stakes educational assessments: Different mixture modeling approaches. *Large-scale Assessments in Education*, *5*(1), 15. https://doi.org/10.1186/s40536-017-0049-3

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 Results (Volume III): What School Life Means for Students' Lives*. OECD. https://doi.org/10.1787/acd78851-en

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion [Place: US Publisher: American Psychological Association]. *Journal of Educational Psychology*, *102*(3), 531–549. https://doi.org/10.1037/a0019243

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education*, *2*(1), 5. https://doi.org/10.1186/s40536-014-0005-4

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

Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. https://doi.org/10.1111/j.1745-3992.2010.00190.x

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, *7*(1), 5. https://doi.org/10.1186/s40536-019-0073-6

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

A. Appendix

In Table A.1, A.4 and A.7 the percentage correct responses of Dutch students are shown for the mathematics, reading and science domain. The numbers are computed per cluster position to provide insight into the development of the students' test performance over the course of the PISA assessment. The tables are moreover split by gender, educational program, and by the PISA indices for socio-economic (ESCS) and immigration status (IMMIG). For the reading domain, only data from 2012 and 2015 are shown as the multistage adaptive test (MSAT) design deployed in 2018 did not allow to allocate responses to the four investigated cluster positions. Furthermore, the group-average percentages for two types of missing response data are shown. Table A.2, A.5 and A.8 contain the percentages item-level non-response of Dutch students per cluster position for the mathematics, reading and science domain. The corresponding percentages of response data coded as *not reached* are shown in Table A.3, A.6 and A.9.

Year	By	Group	1	2	3	4
2012	Gender	Female	54.0	54.3	51.0	49.6
	Gender	Male	56.9	56.5	54.3	51.4
	Educational program	VMBO BB	26.9	26.2	23.7	22.8
	Educational program	VMBO KB	38.4	36.6	33.4	30.7
	Educational program	VMBO GT	49.6	48.7	45.6	42.1
	Educational program	HAVO	62.2	62.1	60.2	57.7
	Educational program	VWO	74.2	75.9	74.3	72.5
	ESCS (quantile)	1	48.2	47.3	45.1	43.0
	ESCS (quantile)	2	51.7	52.0	48.6	45.5
	ESCS (quantile)	3	55.3	54.9	52.4	50.2
	ESCS (quantile)	4	57.6	59.5	56.0	54.1
	ESCS (quantile)	5	65.1	63.3	61.4	60.0
	Immigration status	Native	56.4	56.8	53.9	51.9
	Immigration status	Second-generation	47.5	42.7	42.9	38.4
	Immigration status	First-generation	47.5	46.1	44.0	40.4
2015	Gender	Female	51.6	51.8	49.2	48.2
	Gender	Male	53.5	52.7	50.8	48.1
	Educational program	VMBO BB	26.5	25.0	24.0	19.6
	Educational program	VMBO KB	34.0	35.3	32.8	29.2
	Educational program	VMBO GT	47.8	45.5	44.7	42.2
	Educational program	HAVO	59.2	59.2	58.5	58.0
	Educational program	VWO	71.8	73.5	68.8	68.8
	ESCS (quantile)	1	44.3	44.2	42.0	40.2
	ESCS (quantile)	2	48.8	49.4	45.8	43.9
	ESCS (quantile)	3	50.7	49.4	49.7	46.1
	ESCS (quantile)	4	54.4	54.6	55.1	52.0
	ESCS (quantile)	5	63.6	63.1	57.7	59.0
	Immigration status	Native	53.5	53.4	51.0	48.9
	Immigration status	Second-generation	43.2	43.6	42.9	41.8
	Immigration status	First-generation	39.1	33.8	46.5	45.6
2018	Gender	Female	55.0	54.9	52.2	53.1
	Gender	Male	56.0	56.4	52.3	52.7
	Educational program	VMBO BB	28.6	26.8	27.4	24.3
	Educational program	VMBO KB	33.7	33.0	28.5	29.5
	Educational program	VMBO GT	46.8	47.5	42.7	43.2
	Educational program	HAVO	59.9	59.4	56.7	57.4
	Educational program	VWO	73.2	74.1	71.7	72.9
	ESCS (quantile)	1	47.1	46.2	45.6	45.6
	ESCS (quantile)	2	51.6	50.5	46.4	47.4
	ESCS (quantile)	3	55.3	55.6	52.5	52.1
	ESCS (quantile)	4	57.8	59.0	57.0	57.3
	ESCS (quantile)	5	65.3	66.2	60.6	63.0
	Immigration status	Native	56.8	57.7	53.4	54.3
	Immigration status	Second-generation	47.3	42.7	42.8	40.2
	Immigration status	First-generation	47.5	40.2	44.3	38.7

Table A.1: Mathematics: correct (%) responses in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	2.5	2.9	3.5	3.8
	Gender	Male	2.6	2.9	3.6	4.1
	Educational program	VMBO BB	7.2	7.4	11.5	10.1
	Educational program	VMBO KB	5.0	5.9	5.8	7.8
	Educational program	VMBO GT	2.5	3.2	3.6	4.5
	Educational program	HAVO	1.4	1.6	1.9	2.2
	Educational program	VWO	0.9	0.9	0.9	1.0
	ESCS (quantile)	1	3.4	3.8	4.4	5.2
	ESCS (quantile)	2	2.9	3.3	4.4	4.5
	ESCS (quantile)	3	2.6	2.9	3.1	4.1
	ESCS (quantile)	4	2.4	2.2	3.5	3.0
	ESCS (quantile)	5	1.5	2.4	2.2	2.9
	Immigration status	Native	2.4	2.7	3.4	3.6
	Immigration status	Second-generation	3.9	4.5	4.2	6.8
	Immigration status	First-generation	4.0	4.8	5.8	8.4
2015	Gender	Female	2.0	2.9	4.1	3.9
	Gender	Male	2.2	3.0	3.0	4.2
	Educational program	VMBO BB	5.8	5.9	8.2	8.8
	Educational program	VMBO KB	4.0	6.6	7.3	8.0
	Educational program	VMBO GT	2.1	2.8	3.8	4.4
	Educational program	HAVO	1.1	1.8	1.8	1.6
	Educational program	VWO	0.6	1.1	1.2	2.0
	ESCS (quantile)	1	2.9	4.6	4.9	5.0
	ESCS (quantile)	2	2.8	3.6	3.4	4.7
	ESCS (quantile)	3	1.7	2.7	3.4	4.6
	ESCS (quantile)	4	1.9	2.5	3.3	3.2
	ESCS (quantile)	5	1.1	1.3	2.9	2.6
	Immigration status	Native	2.0	2.8	3.5	3.8
	Immigration status	Second-generation	2.6	3.6	4.5	5.8
	Immigration status	First-generation	0.7	2.2	2.1	5.4
2018	Gender	Female	2.4	2.8	3.5	3.6
	Gender	Male	2.7	3.3	3.9	4.4
	Educational program	VMBO BB	7.7	8.5	7.1	7.9
	Educational program	VMBO KB	5.3	7.3	9.7	10.1
	Educational program	VMBO GT	2.9	3.9	3.7	4.2
	Educational program	HAVO	1.8	1.6	2.7	3.3
	Educational program	VWO	1.2	1.3	1.6	1.4
	ESCS (quantile)	1	3.5	3.8	4.9	5.7
	ESCS (quantile)	2	2.5	3.6	4.7	4.2
	ESCS (quantile)	3	3.2	3.8	2.9	3.8
	ESCS (quantile)	4	2.3	2.2	2.6	2.7
	ESCS (quantile)	5	1.3	2.1	3.2	3.5
	Immigration status	Native	2.5	2.8	3.5	3.7
	Immigration status	Second-generation	2.4	4.3	5.1	7.4
	Immigration status	First-generation	6.1	6.0	8.6	7.0

Table A.2: Mathematics: non-response (%) on item-level in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	0.0	0.1	0.2	0.7
	Gender	Male	0.0	0.0	0.1	0.5
	Educational program	VMBO BB	0.0	0.0	0.0	1.4
	Educational program	VMBO KB	0.0	0.0	0.3	0.4
	Educational program	VMBO GT	0.0	0.1	0.3	1.2
	Educational program	HAVO	0.0	0.0	0.0	0.3
	Educational program	VWO	0.0	0.0	0.0	0.2
	ESCS (quantile)	1	0.0	0.0	0.1	0.7
	ESCS (quantile)	2	0.0	0.0	0.2	0.4
	ESCS (quantile)	3	0.0	0.0	0.1	0.7
	ESCS (quantile)	4	0.0	0.2	0.4	0.5
	ESCS (quantile)	5	0.0	0.0	0.0	0.8
	Immigration status	Native	0.0	0.0	0.2	0.7
	Immigration status	Second-generation	0.0	0.0	0.1	0.2
	Immigration status	First-generation	0.0	0.0	0.0	0.7
2015	Gender	Female	0.5	0.7	0.6	1.3
	Gender	Male	0.8	0.9	0.7	0.4
	Educational program	VMBO BB	3.8	4.3	2.6	2.2
	Educational program	VMBO KB	0.7	0.8	1.6	3.2
	Educational program	VMBO GT	0.2	0.6	0.4	0.5
	Educational program	HAVO	0.0	0.2	0.3	0.1
	Educational program	VWO	0.6	0.5	0.1	0.3
	ESCS (quantile)	1	0.9	1.7	1.1	1.4
	ESCS (quantile)	2	1.1	0.8	1.2	1.4
	ESCS (quantile)	3	0.1	0.5	0.3	1.1
	ESCS (quantile)	4	0.4	0.4	0.2	0.4
	ESCS (quantile)	5	0.6	0.5	0.4	0.3
	Immigration status	Native	0.4	0.5	0.4	0.6
	Immigration status	Second-generation	2.1	2.1	1.7	2.0
	Immigration status	First-generation	8.7	10.8	2.8	8.1
2018	Gender	Female	0.5	1.3	0.6	0.8
	Gender	Male	0.4	0.9	0.7	0.5
	Educational program	VMBO BB	2.4	6.4	1.2	0.6
	Educational program	VMBO KB	1.6	1.9	1.2	0.7
	Educational program	VMBO GT	0.2	0.5	0.6	0.4
	Educational program	HAVO	0.3	0.9	0.7	0.6
	Educational program	VWO	0.0	1.1	0.4	0.8
	ESCS (quantile)	1	0.7	1.5	0.8	0.5
	ESCS (quantile)	2	0.4	1.3	0.6	0.5
	ESCS (quantile)	3	0.4	0.5	0.5	0.2
	ESCS (quantile)	4	0.1	0.6	0.2	0.7
	ESCS (quantile)	5	0.5	1.8	1.2	1.3
	Immigration status	Native	0.3	1.0	0.6	0.6
	Immigration status	Second-generation	0.7	1.4	1.7	1.0
	Immigration status	First-generation	0.0	4.0	2.9	0.0

Table A.3: Mathematics: not reached (%) on item-level in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	67.8	62.7	61.9	59.9
	Gender	Male	63.0	60.9	57.3	53.6
	Educational program	VMBO BB	36.0	36.4	32.8	28.0
	Educational program	VMBO KB	48.9	44.4	40.4	35.2
	Educational program	VMBO GT	60.8	56.3	53.6	50.5
	Educational program	HAVO	71.4	70.1	66.9	62.6
	Educational program	VWO	83.2	80.3	78.6	78.3
	ESCS (quantile)	1	55.8	52.1	55.3	48.4
	ESCS (quantile)	2	64.2	58.4	58.9	55.6
	ESCS (quantile)	3	65.4	59.3	54.8	57.3
	ESCS (quantile)	4	70.0	67.3	63.7	59.8
	ESCS (quantile)	5	71.8	71.9	65.5	61.4
	Immigration status	Native	66.5	62.5	60.5	57.7
	Immigration status	Second-generation	54.9	57.6	49.4	49.6
	Immigration status	First-generation	66.8	50.6	55.6	52.3
2015	Gender	Female	70.0	67.4	65.5	63.9
	Gender	Male	65.7	63.5	61.4	57.5
	Educational program	VMBO BB	39.4	38.9	35.6	31.4
	Educational program	VMBO KB	53.0	49.3	44.3	41.1
	Educational program	VMBO GT	65.3	60.7	57.7	53.0
	Educational program	HAVO	74.9	73.6	72.0	71.3
	Educational program	VWO	82.1	81.2	81.9	80.2
	ESCS (quantile)	1	59.9	57.5	55.6	52.7
	ESCS (quantile)	2	65.9	60.8	57.6	54.9
	ESCS (quantile)	3	66.2	64.8	62.0	60.9
	ESCS (quantile)	4	70.9	68.9	67.6	64.8
	ESCS (quantile)	5	76.7	75.2	74.3	70.3
	Immigration status	Native	68.4	66.2	64.7	62.0
	Immigration status	Second-generation	63.6	60.3	54.9	53.4
	Immigration status	First-generation	62.3	58.6	52.6	46.8

Table A.4: Reading: correct (%) responses in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	1.3	1.4	1.8	2.0
	Gender	Male	1.9	1.8	2.3	3.1
	Educational program	VMBO BB	6.1	5.0	5.6	7.9
	Educational program	VMBO KB	2.5	3.1	3.9	4.7
	Educational program	VMBO GT	1.4	1.2	2.5	2.6
	Educational program	HAVO	0.9	0.8	1.0	1.4
	Educational program	VWO	0.4	0.4	0.4	0.8
	ESCS (quantile)	1	2.2	2.0	2.2	4.3
	ESCS (quantile)	2	1.8	1.6	1.9	2.3
	ESCS (quantile)	3	1.2	2.4	2.5	2.2
	ESCS (quantile)	4	1.2	1.2	2.0	2.2
	ESCS (quantile)	5	1.5	0.7	1.8	1.9
	Immigration status	Native	1.6	1.5	1.9	2.4
	Immigration status	Second-generation	1.7	1.8	3.6	2.9
	Immigration status	First-generation	1.3	3.3	3.2	4.7
2015	Gender	Female	1.0	0.9	1.4	1.9
	Gender	Male	1.3	1.2	1.4	2.2
	Educational program	VMBO BB	5.0	3.5	6.1	5.2
	Educational program	VMBO KB	2.5	2.0	3.2	5.1
	Educational program	VMBO GT	0.9	1.0	1.5	2.5
	Educational program	HAVO	0.3	0.4	0.3	0.6
	Educational program	VWO	0.2	0.3	0.2	0.3
	ESCS (quantile)	1	2.4	1.3	2.9	3.6
	ESCS (quantile)	2	1.2	1.4	1.2	2.2
	ESCS (quantile)	3	0.8	1.2	1.7	1.7
	ESCS (quantile)	4	1.0	0.6	0.9	1.6
	ESCS (quantile)	5	0.4	0.7	0.7	1.4
	Immigration status	Native	1.1	1.0	1.4	1.9
	Immigration status	Second-generation	1.7	1.5	2.4	3.7
	Immigration status	First-generation	4.3	3.7	0.9	2.3

Table A.5: Reading: non-response (%) on item-level in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	0.0	0.1	0.2	1.3
	Gender	Male	0.0	0.0	0.1	1.1
	Educational program	VMBO BB	0.0	0.0	0.4	0.9
	Educational program	VMBO KB	0.0	0.0	0.0	2.6
	Educational program	VMBO GT	0.0	0.2	0.4	1.1
	Educational program	HAVO	0.0	0.0	0.0	0.7
	Educational program	VWO	0.0	0.0	0.0	1.4
	ESCS (quantile)	1	0.0	0.0	0.2	1.5
	ESCS (quantile)	2	0.0	0.0	0.0	0.7
	ESCS (quantile)	3	0.0	0.3	0.2	1.4
	ESCS (quantile)	4	0.0	0.0	0.0	1.7
	ESCS (quantile)	5	0.0	0.0	0.4	0.9
	Immigration status	Native	0.0	0.0	0.2	1.2
	Immigration status	Second-generation	0.0	0.7	0.0	1.8
	Immigration status	First-generation	0.0	0.0	0.0	1.9
2015	Gender	Female	0.0	0.1	0.1	0.2
	Gender	Male	0.5	0.7	0.2	0.4
	Educational program	VMBO BB	2.3	2.4	0.9	2.2
	Educational program	VMBO KB	0.4	0.1	0.2	0.2
	Educational program	VMBO GT	0.0	0.3	0.1	0.3
	Educational program	HAVO	0.0	0.1	0.0	0.1
	Educational program	VWO	0.0	0.3	0.0	0.0
	ESCS (quantile)	1	0.4	0.1	0.0	0.3
	ESCS (quantile)	2	0.3	0.4	0.2	0.2
	ESCS (quantile)	3	0.6	1.1	0.3	0.8
	ESCS (quantile)	4	0.0	0.5	0.0	0.0
	ESCS (quantile)	5	0.0	0.2	0.1	0.1
	Immigration status	Native	0.2	0.5	0.1	0.1
	Immigration status	Second-generation	0.8	0.0	0.8	2.0
	Immigration status	First-generation	0.0	0.9	0.0	0.0

Table A.6: Reading: not reached (%) on item-level in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	58.9	59.3	58.7	53.2
	Gender	Male	60.7	61.6	57.4	55.6
	Educational program	VMBO BB	36.4	35.1	29.6	28.2
	Educational program	VMBO KB	44.5	43.2	39.9	33.1
	Educational program	VMBO GT	56.1	54.9	52.1	48.8
	Educational program	HAVO	66.7	68.2	63.7	63.2
	Educational program	VWO	75.8	77.2	77.8	74.7
	ESCS (quantile)	1	52.9	51.5	49.4	43.9
	ESCS (quantile)	2	55.6	57.1	54.3	51.1
	ESCS (quantile)	3	58.2	60.4	59.0	52.3
	ESCS (quantile)	4	64.5	63.8	61.7	60.2
	ESCS (quantile)	5	67.3	70.4	66.1	65.9
	Immigration status	Native	60.9	61.7	59.6	55.4
	Immigration status	Second-generation	51.7	48.6	44.5	43.2
	Immigration status	First-generation	51.1	57.6	49.4	49.4
2015	Gender	Female	54.3	53.5	53.7	51.0
	Gender	Male	56.9	54.3	54.6	52.3
	Educational program	VMBO BB	26.4	25.3	24.6	24.8
	Educational program	VMBO KB	37.0	33.8	35.9	32.0
	Educational program	VMBO GT	51.2	48.5	47.3	45.1
	Educational program	HAVO	64.1	62.4	62.4	58.8
	Educational program	VWO	74.2	74.3	74.3	73.0
	ESCS (quantile)	1	45.8	43.7	44.9	42.2
	ESCS (quantile)	2	52.6	48.5	50.0	47.7
	ESCS (quantile)	3	54.5	54.1	52.9	48.8
	ESCS (quantile)	4	60.7	59.4	57.3	56.7
	ESCS (quantile)	5	64.3	63.9	65.5	62.6
	Immigration status	Native	56.9	55.1	55.6	53.1
	Immigration status	Second-generation	45.4	44.1	43.5	40.5
	Immigration status	First-generation	44.9	51.0	37.7	33.5
2018	Gender	Female	59.4	53.9	56.5	52.6
	Gender	Male	60.1	54.1	55.7	52.2
	Educational program	VMBO BB	23.7	20.1	30.5	27.4
	Educational program	VMBO KB	38.9	32.7	33.4	27.2
	Educational program	VMBO GT	52.4	47.2	47.1	42.3
	Educational program	HAVO	65.6	59.4	61.1	57.7
	Educational program	VWO	76.7	70.7	74.0	72.3
	ESCS (quantile)	1	51.8	47.1	47.0	40.3
	ESCS (quantile)	2	55.2	50.2	52.3	50.0
	ESCS (quantile)	3	61.1	53.9	54.9	52.1
	ESCS (quantile)	4	62.5	56.8	58.9	55.4
	ESCS (quantile)	5	67.8	61.8	67.8	64.6
	Immigration status	Native	61.0	55.7	58.0	54.5
	Immigration status	Second-generation	48.1	40.2	41.2	36.5
	Immigration status	First-generation	50.2	35.9	44.6	35.0

Table A.7: Science: correct (%) responses in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	0.8	1.2	1.5	2.3
	Gender	Male	0.9	0.8	1.4	1.5
	Educational program	VMBO BB	3.1	4.0	4.5	6.0
	Educational program	VMBO KB	1.6	1.7	3.1	3.5
	Educational program	VMBO GT	0.5	1.0	1.5	2.0
	Educational program	HAVO	0.4	0.5	0.7	0.7
	Educational program	VWO	0.4	0.2	0.1	0.5
	ESCS (quantile)	1	1.3	1.4	1.9	2.9
	ESCS (quantile)	2	1.3	1.3	1.9	2.6
	ESCS (quantile)	3	0.7	1.0	1.4	1.4
	ESCS (quantile)	4	0.5	0.8	0.8	1.1
	ESCS (quantile)	5	0.6	0.7	1.0	1.5
	Immigration status	Native	0.8	0.9	1.4	1.8
	Immigration status	Second-generation	0.9	1.9	1.4	3.3
	Immigration status	First-generation	2.4	1.3	2.3	4.2
2015	Gender	Female	1.1	1.1	1.2	1.5
	Gender	Male	1.3	1.2	1.3	1.5
	Educational program	VMBO BB	2.7	2.9	4.1	4.8
	Educational program	VMBO KB	3.1	2.9	2.7	3.2
	Educational program	VMBO GT	1.1	1.3	1.2	1.2
	Educational program	HAVO	0.5	0.4	0.6	0.8
	Educational program	VWO	0.2	0.2	0.3	0.4
	ESCS (quantile)	1	1.9	1.7	1.9	2.7
	ESCS (quantile)	2	1.2	1.8	1.6	1.5
	ESCS (quantile)	3	0.9	1.0	1.2	1.4
	ESCS (quantile)	4	0.8	0.6	0.9	1.5
	ESCS (quantile)	5	1.1	0.7	0.8	0.6
	Immigration status	Native	1.1	1.1	1.2	1.4
	Immigration status	Second-generation	2.1	2.0	1.5	2.3
	Immigration status	First-generation	1.1	1.7	3.6	4.8
2018	Gender	Female	0.8	1.3	1.4	1.6
	Gender	Male	0.8	1.6	1.3	1.7
	Educational program	VMBO BB	2.0	5.5	2.8	3.4
	Educational program	VMBO KB	2.3	4.4	3.0	4.6
	Educational program	VMBO GT	0.9	1.5	1.6	2.3
	Educational program	HAVO	0.4	0.7	1.2	0.9
	Educational program	VWO	0.1	0.2	0.2	0.3
	ESCS (quantile)	1	1.1	2.4	1.8	2.2
	ESCS (quantile)	2	1.0	1.9	1.8	2.0
	ESCS (quantile)	3	0.6	1.1	1.6	2.4
	ESCS (quantile)	4	0.7	1.2	1.1	1.2
	ESCS (quantile)	5	0.4	0.6	0.4	0.5
	Immigration status	Native	0.8	1.2	1.2	1.2
	Immigration status	Second-generation	0.9	3.2	1.7	4.8
	Immigration status	First-generation	0.0	2.4	3.7	8.5

Table A.8: Science: non-response (%) on item-level in NL by cluster position.

Year	By	Group	1	2	3	4
2012	Gender	Female	0.0	0.0	0.2	1.0
	Gender	Male	0.0	0.0	0.1	0.6
	Educational program	VMBO BB	0.0	0.0	0.0	0.0
	Educational program	VMBO KB	0.0	0.0	0.0	0.8
	Educational program	VMBO GT	0.0	0.0	0.3	1.1
	Educational program	HAVO	0.0	0.0	0.3	0.9
	Educational program	VWO	0.0	0.0	0.0	0.8
	ESCS (quantile)	1	0.0	0.0	0.0	0.8
	ESCS (quantile)	2	0.0	0.0	0.0	0.9
	ESCS (quantile)	3	0.0	0.0	0.0	1.0
	ESCS (quantile)	4	0.0	0.0	0.5	0.7
	ESCS (quantile)	5	0.0	0.0	0.4	0.6
	Immigration status	Native	0.0	0.0	0.2	0.6
	Immigration status	Second-generation	0.0	0.0	0.0	3.1
	Immigration status	First-generation	0.0	0.0	0.0	2.0
2015	Gender	Female	0.4	0.7	0.3	0.6
	Gender	Male	0.5	0.8	0.5	0.8
	Educational program	VMBO BB	1.0	2.2	1.2	2.5
	Educational program	VMBO KB	1.0	1.5	0.7	0.8
	Educational program	VMBO GT	0.3	0.6	0.3	0.7
	Educational program	HAVO	0.2	0.4	0.2	0.5
	Educational program	VWO	0.4	0.4	0.3	0.4
	ESCS (quantile)	1	1.1	1.7	0.7	1.1
	ESCS (quantile)	2	0.4	0.9	0.5	0.9
	ESCS (quantile)	3	0.4	0.4	0.2	0.5
	ESCS (quantile)	4	0.2	0.3	0.3	0.4
	ESCS (quantile)	5	0.2	0.4	0.4	0.7
	Immigration status	Native	0.3	0.5	0.4	0.7
	Immigration status	Second-generation	2.3	3.0	0.6	0.9
	Immigration status	First-generation	0.2	0.0	1.0	3.4
2018	Gender	Female	0.1	3.8	0.8	3.1
	Gender	Male	0.5	3.6	0.5	1.5
	Educational program	VMBO BB	3.9	9.7	1.8	4.9
	Educational program	VMBO KB	0.5	1.6	1.7	3.6
	Educational program	VMBO GT	0.4	2.2	1.1	1.4
	Educational program	HAVO	0.1	4.1	0.2	2.7
	Educational program	VWO	0.0	5.3	0.0	2.0
	ESCS (quantile)	1	1.0	3.9	1.5	3.5
	ESCS (quantile)	2	0.1	2.6	0.5	2.7
	ESCS (quantile)	3	0.2	3.6	1.0	1.5
	ESCS (quantile)	4	0.2	4.1	0.2	2.0
	ESCS (quantile)	5	0.1	4.1	0.2	1.9
	Immigration status	Native	0.3	3.2	0.4	1.9
	Immigration status	Second-generation	0.5	6.9	0.6	3.4
	Immigration status	First-generation	0.0	15.3	8.6	14.8

Table A.9: Science: not reached (%) on item-level in NL by cluster position.

List of Figures

List of Tables