CPB Netherlands Bureau for
Economic Policy Analysis

A look at the Dutch position in international student assessments

International student assessments indicate a deterioration in Dutch educational performance.
In this publication, we examine whether the methods used in these assessments influence the results. We note that the position of the Netherlands is sensitive to the influx of new countries and that methodological problems can affect the scores.
It is unknown whether the totality of these problems has had a favorable or unfavorable effect on the position of the Netherlands.

To better interpret these international student assessments and make them more useful for policymaking, it would be helpful if the coordinating organizations analyze the impact of methodological choices more extensively and make it possible to link their results to national tests.

CPB - December 2022

Paul Verstraten,
Annikka Lemmens,
Marielle Non

Summary

Concerns have arisen in recent years about the results of international student assessments, which seem to indicate that the level of basic skills in the Netherlands is falling. In the 3-yearly assessment of the *Programme for International Student Assessment* (PISA) among 15-year-old secondary school students, the score of the Netherlands in the areas of reading and science is deteriorating since 2015. The PISA mathematics score has been falling since 2003, but showed a slight recovery in 2018 (the last reference year). The position of the Netherlands at PISA is also deteriorating relative to other countries. Research by the *Trends in International Mathematics and Science Study* (TIMSS), which focuses on 10-year-olds, shows a slightly decreasing score for the Netherlands in science and mathematics since 1995. The position of the Netherlands in this TIMSS ranking is declining sharply compared to other countries. The *Progress in International Reading Literacy Study* (PIRLS), on the other hand, shows that the reading skills of 10-year-olds have remained fairly stable between 2001 and 2016. Partly due to these assessments, more attention is currently being paid to the level of basic skills in the Netherlands.

This study examines to what extent international student assessments are an accurate indicator of educational performance and concludes that various problems can influence the score and position of the Netherlands. For example, it appears that alterations of the assessment mode, such as a switch from paper-based to computer-based testing, work out more favorably for some countries than for others. This makes comparison of the scores over time and between countries more difficult. The coordinating organizations do use statistical methods to correct for differences in the assessment mode, but they do not always succeed in doing so completely. In addition, the statistical methods themselves are regularly changed and improved. Such adjustments improve the reliability of the scores in the new reference years, but make comparisons with previous reference years more difficult. We also note that the coordinating organizations regularly review the assessment framework. Over the years, for example, the meaning of 'good reading' has changed. This has advantages, but does not benefit the comparability of the results over time. Differences between countries in the motivation with which students take the test and flaws in the sample can also affect reliability and comparability. The sum of these problems complicate a fair comparison of the scores between countries or between reference years, so that differences in scores do not always equate to differences in the knowledge and skills of pupils.

The literature shows that these problems can influence the position of the Netherlands both positively and negatively, although it is not possible to capture the total effect of the problems in one figure. A number of studies show that the position of the Netherlands in PISA mathematics in 2015 and PISA reading in 2015 and 2018 has been adversely affected by changes in the assessment mode, the assessment framework and the statistical methods. Other research has looked at the influence of motivation on the score of the Netherlands. This shows that Dutch students are on average more serious than in other countries, which positively influenced the score of the Netherlands in PISA science in 2015. However, differences in motivation are not equal to differences in knowledge or skills. Finally, there are indications that non-response may influence the position of the Netherlands, although the effect of this on the Dutch position is unclear. Because the literature is highly fragmented and often incomplete, it is not possible to quantify the influence of all these factors on the score and ranking of the Netherlands.

The international student assessments could implement a number of changes that make the above-mentioned problems more transparent and contribute to solutions. Firstly, it would be useful if the PISA data could be linked to national tests (such as the final test in primary education and the final exam in secondary education) and the student tracking system (LVS). This is currently not possible, which hinders for example research into the quality of the sample. Secondly, it would be good if the international assessments

were more transparent about the consequences of important choices that are inevitably somewhat arbitrary. This includes assumptions in the statistical modelling. Thirdly, the international assessments could do more to provide insight into the influence of changes in the assessment framework.

1 Introduction

International student assessments monitor the cognitive skills of students from different countries. The most comprehensive international student assessments are the *Programme for International Student Assessment* (PISA), the *Trends in International Mathematics and Science Study* (TIMSS) and the *Progress in International Reading Literacy Study* (PIRLS). PISA is conducted every three years by the *Organisation for Economic Co-operation and Development* (OECD) and TIMSS and PIRLS are released by the *International Association for the Evaluation of Educational Achievement* (IEA) every four and five years respectively. These studies focus on different disciplines and age groups. PIRLS focuses on reading skills among 10-year-olds and TIMSS focuses on math and science among the same age group.[1] PISA covers the domains of reading, mathematics and science among 15-year-olds.

The results of international assessments are often used by countries to implement major policy changes in education. A strong policy response triggered by unexpectedly poor PISA results is referred to in the literature as the "PISA shock" (Breakspear, 2014). Baird et al. (2011) cite France as an example, where, in response to the PISA 2009 results, it was announced that primary education would focus more on basic skills and personalized education. Another example that shows the influence of international assessments is the case of Norway. When Norway was found to be performing below the PISA average in 2000 and 2003, reforms were implemented targeting the curriculum and testing. According to Wiseman (2013), Germany and Japan have also implemented major policy changes in the past due to disappointing results in international assessments.

In recent years, concerns have also arisen in the Netherlands about the deteriorating position of the Netherlands in international student assessments. The most recent scores of the Netherlands, at the end of the 2010s, are lower in all international assessments than the score recorded by the Netherlands in the first reference year, in the mid-1990s and early 2000s. In addition to the decrease in the score, the Netherlands also saw a deteriorating position compared to other countries. While two decades ago the Netherlands enjoyed a fairly solid top-10 position in all international assessments, this is now only the case for PISA mathematics and science. Concerns about this development were expressed in, inter alia, 'De Staat van het Onderwijs' (Inspectorate of Education, 2021, p. 98) and the letter to parliament on the Basic Skills Master Plan (OCW, 2022, p. 3). As an aside, critics note that these assessments ignore the importance of non-cognitive skills, such as social skills and perseverance. The results of international student assessments would therefore only be a limited indicator of educational quality (Araujo et al., 2017).

At the request of the Ministry of Education, Culture and Science, the CPB has investigated the extent to which international assessments are an accurate indicator of Dutch educational performance. The analysis consists of three parts. Section 2 analyzes the development of the Netherlands in international student assessments. In section 3 we provide an overview of – what we call – equivalence problems in international assessments, which make it difficult to compare countries with each other and over time. In section 4 we discuss the consequences of these equivalence problems for the position of the Netherlands.

[1] We do not consider TIMSS for secondary education because the Netherlands has not participated in this study since 2008.

2 The position of the Netherlands in international student assessments

In this section we look at the position of the Netherlands in international assessments based on three criteria: the absolute score, the total ranking and a consistent ranking. The absolute score concerns the score points of the Netherlands. The total ranking shows the position of the Netherlands compared to all other participating countries. However, the total ranking is affected by the increase in participants over time. That is why we also show a consistent ranking: a ranking with only those countries that participated in the same reference years as the Netherlands.[2] This ranking is not distorted by the inflow and outflow of participants, but has the disadvantage that only a few countries remain in some rankings.

Reading

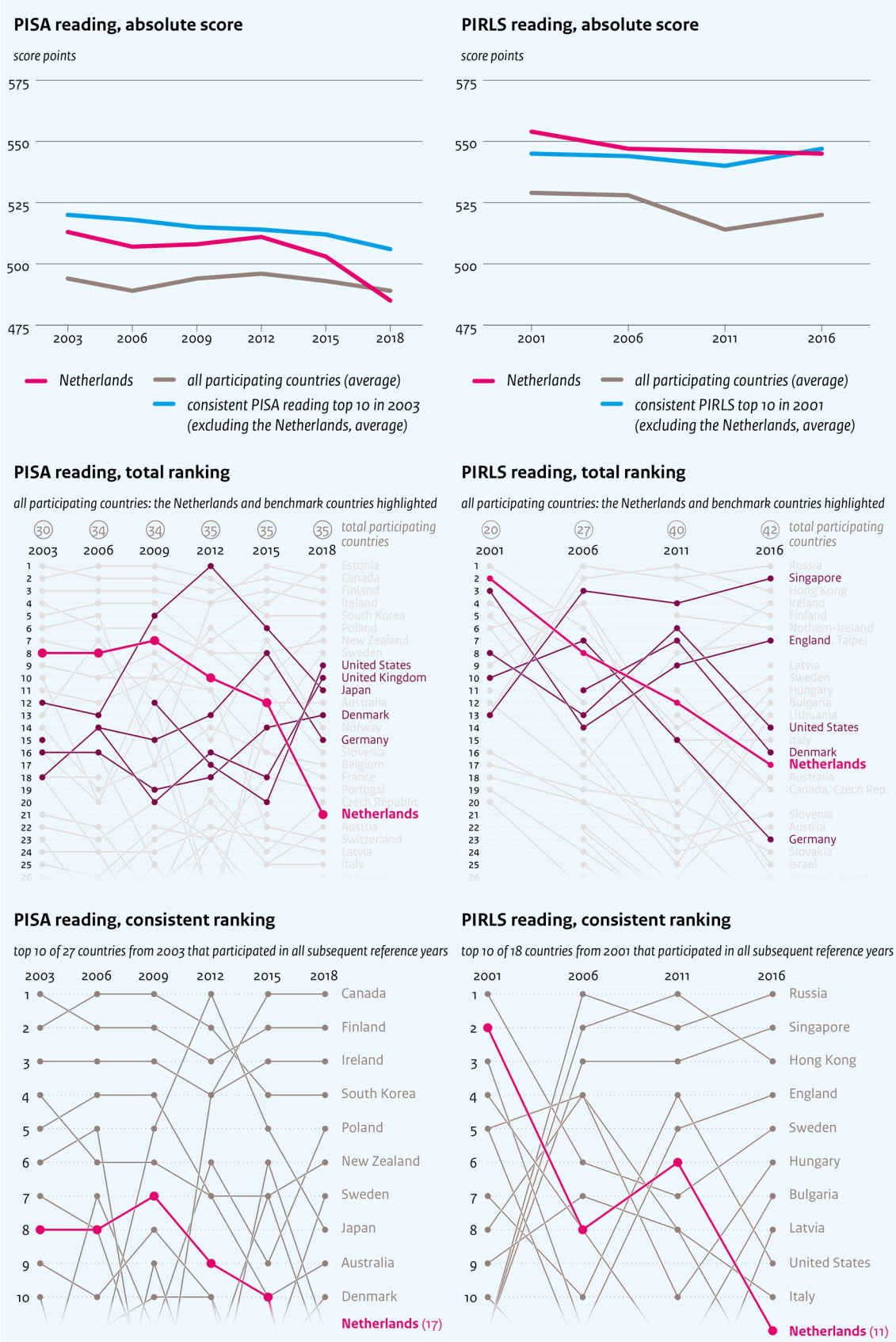
The absolute score of the Netherlands in PISA reading is fairly stable until 2012, but has dropped considerably since then, causing the Netherlands to fall in both the total and the consistent ranking. The left half of figure 2.1 shows the development of the Netherlands in PISA reading. The score of the Netherlands remained at the same level until 2012, but after that it fell by as much as 18 points between 2015 and 2018. The Netherlands therefore scored around the average of all participants in 2018. The top 10 best performing countries show a slight decrease in the absolute score over time. However, the score of the Netherlands has fallen much more sharply in the last two reference years, as a result of which the Netherlands has slipped considerably in the ranking. The figure highlights a number of benchmark countries that are reasonably comparable to the Netherlands in terms of economic development. The Netherlands has also fallen sharply compared to these countries. The fall of the Netherlands in the ranking can only be explained to a very limited extent by the influx of new countries. After all, the consistent ranking also shows a significant drop, from 8th place in 2003 to 17th place in 2018.

Dutch 10-year-olds score relatively stable on the reading test of PIRLS – after a slight decrease between 2001 and 2006 – but the Netherlands is falling further and further on the ranking because other countries are improving their scores and more countries are participating. The right half of figure 2.1 shows the development of the Netherlands at PIRLS. In the first reference years, the Netherlands scores quite high compared to the best performing countries. However, between 2011 and 2016, other countries experienced an improvement in their scores, while Dutch pupils' performance remained stable. As a result, the Netherlands has dropped a number of places in the consistent ranking. The Netherlands has dropped more in the overall ranking, which, in addition to the increase in the score of other countries, can also be explained by an increase in the number of participants.

The decline in reading skills of 15-year-olds measured by PISA between 2012 and 2018 contrasts sharply with the almost constant score of 10-year-olds in PIRLS between 2006 and 2016. These tests largely concern the same cohorts of students. A possible explanation for the difference may be that reading skills developed relatively weakly between the ages of ten and fifteen or that motivation declined sharply. Another explanation may be that the studies do not measure the same thing. For example, PISA 2018 gives more weight to the evaluation and reflection component than PIRLS 2011 and 2016 (IEA, 2009, p. 14; IEA, 2015, p. 14; OECD, 2019a, p. 42). Section 3 takes a closer look at the various aspects that can distort the score in these assessments.

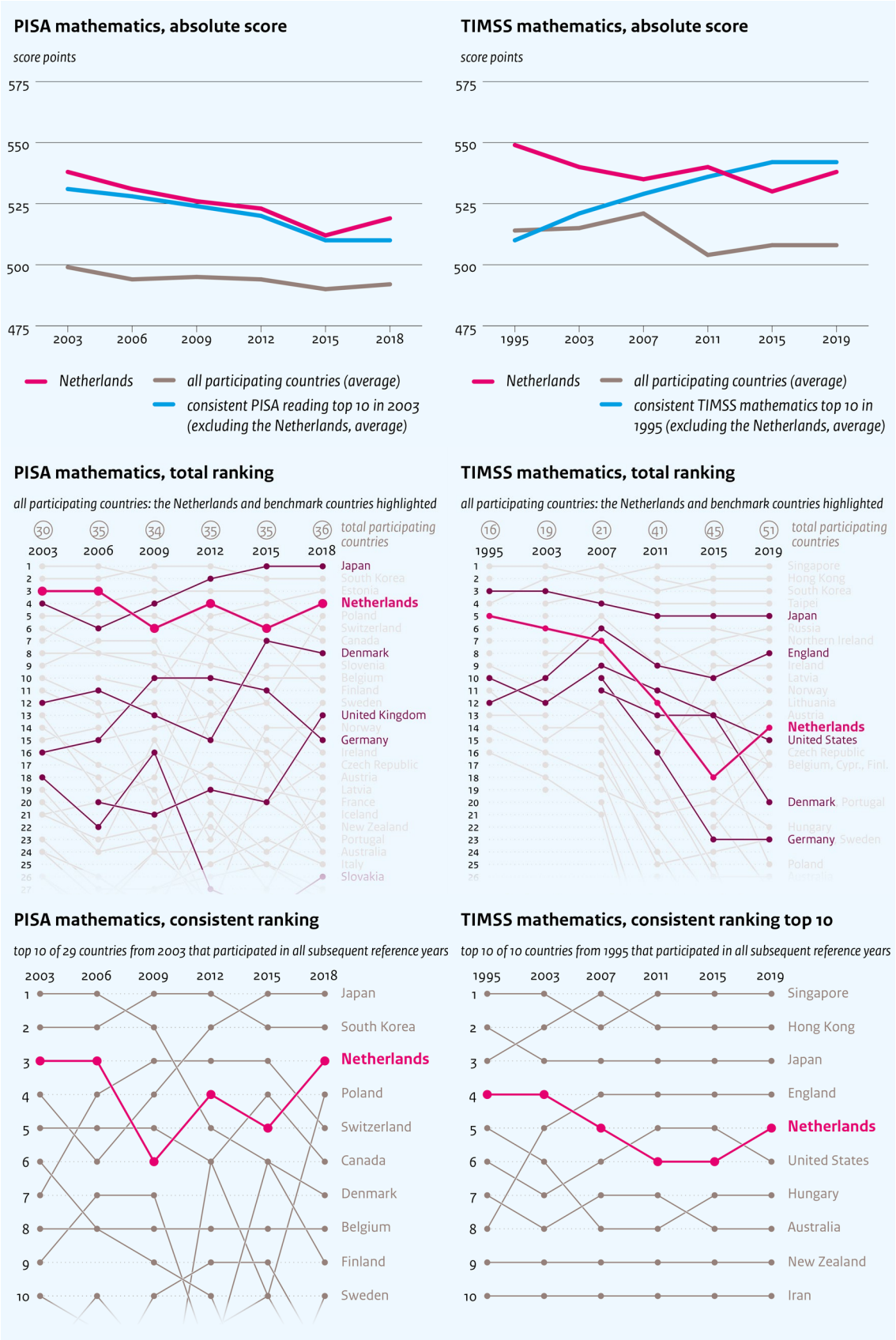
[2] The Netherlands has participated in almost all reference years of PISA, TIMSS and PIRLS. The Netherlands also participated in PISA 2000, but the non-response was too high, so no results were published for the Netherlands. In 1999, the Netherlands only participated in the test for secondary education at TIMSS. We ignore these results.

Figure 2.1 Reading skills in PISA (15-year-olds, left) and PIRLS (10-year-olds, right)



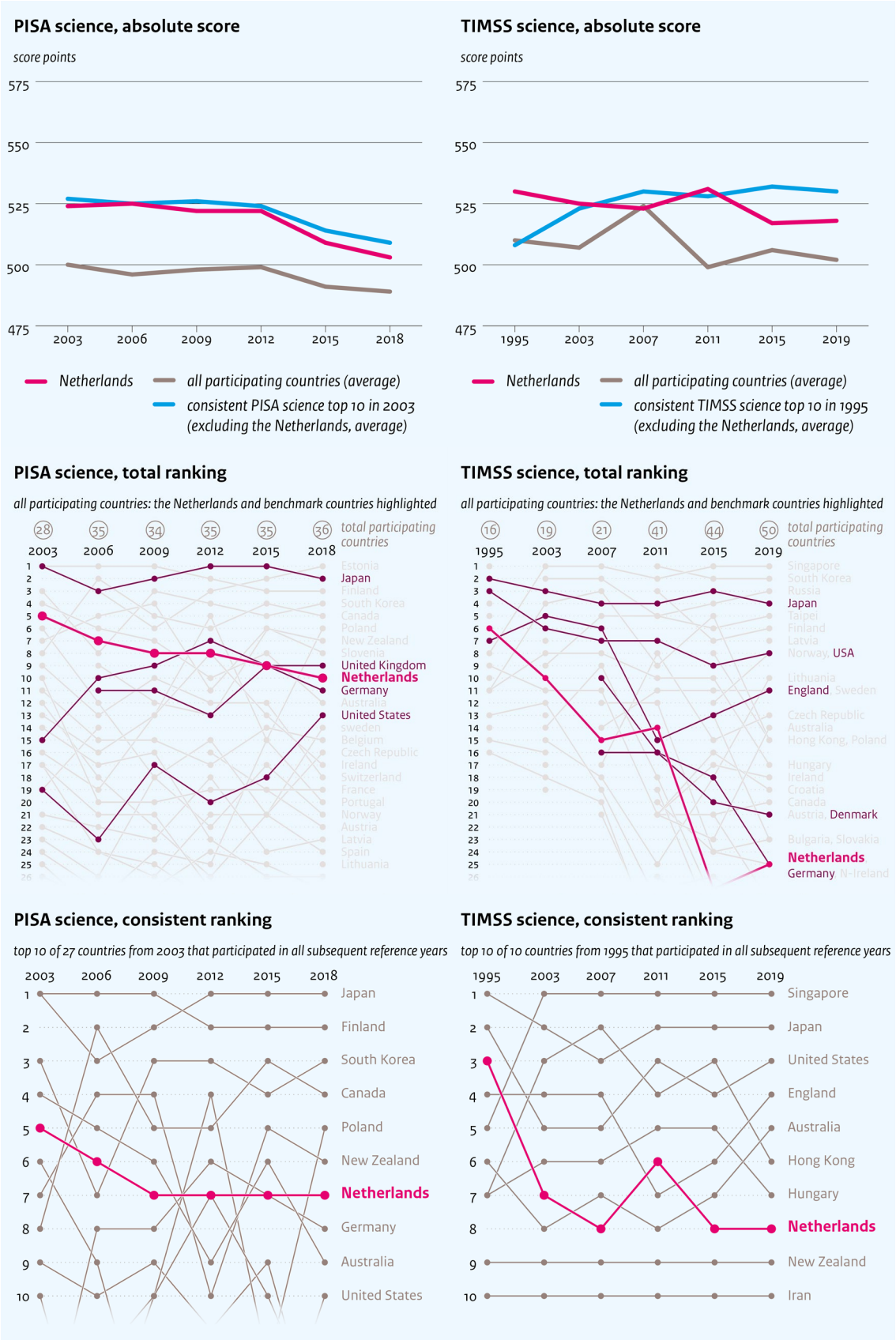
Source: own analysis on the basis of onderwijsincijfers.nl (PISA, left) and the Download Center of the IEA (PIRLS, right).

Figure 2.2 Mathematics in PISA (15-year-olds, left) and TIMSS (10-year-olds, right)



Source: own analysis on the basis of onderwijsincijfers.nl (PISA, left) and the Download Center of the IEA (TIMSS, right).

Figure 2.3 Science at PISA (15-year-olds, left) and TIMSS (10-year-olds, right)



Source: own analysis on the basis of onderwijsincijfers.nl (PISA, left) and the Download Center of the IEA (TIMSS, right).

Mathematics and science

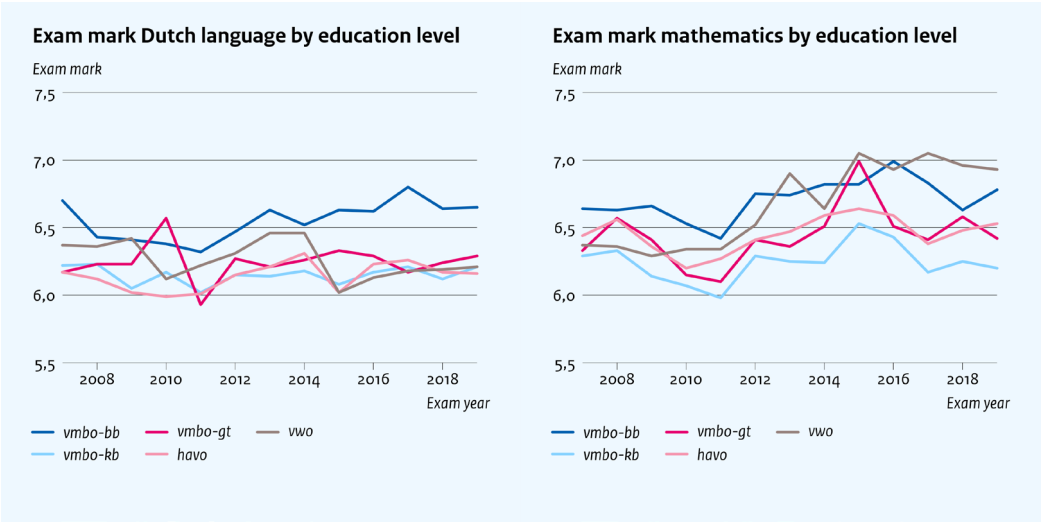
In PISA mathematics and science, the Netherlands shows a downward trend in the absolute score, but the top-10 ranking is maintained. The left half of Figures 2.2 and 2.3 shows the development of the Netherlands in PISA mathematics and science respectively. The absolute score in mathematics has shown a downward trend since the first reference year, although the score rose again slightly in 2018. In science, the score has fallen sharply, particularly in the last two reference years. However, the top 10 well-performing countries show a similar downward trend, as a result of which the Netherlands only drops slightly in both the total and the consistent ranking. The differences between the top countries are small, so that the ranking has an irregular pattern. For example, in 2012 the Netherlands scored one point higher than Australia in science and two points lower than Germany. In both cases the difference was insignificant. In 2015, the Netherlands scored just one point lower than Australia and equal to Germany. A small change in the absolute score can therefore lead to an increase or decrease of one or more places in the ranking.

The absolute score of the Netherlands in TIMSS mathematics and science has fallen slightly, but the Netherlands has fallen considerably in the ranking because the top 10 countries have started to perform better and there has been a substantial influx of participants over time. The right half of Figures 2.2 and 2.3 shows the development of the Netherlands in TIMSS mathematics and science respectively. The results in 1995 and 2019 are subject to larger-than-usual uncertainty because the Netherlands did not reach the required number of participating schools in those reference years. The absolute score fluctuates somewhat, but shows a slight decrease over the entire period. Other well-performing countries do not show such a decline and, in combination with an increase in the number of participating countries, this means that the Netherlands has fallen sharply in the total ranking: in mathematics from 5th place in 1995 to 14th place in 2019 and in science from 6th place in 1995 to 25th place in 2019. The consistent ranking shows a more modest decline, but this picture may be distorted because there are few countries that have participated in all reference years.

National tests show a different picture, but are difficult to compare

National tests – such as the final tests in primary education, central final exams in secondary education, and national surveys – partly show a different picture of basic skills in the Netherlands. National figures on proficiency in reading and mathematics are scarce in the Netherlands (Dutch Education Council, 2022), but the figures that are available do not always show the same picture as international student assessments. For example, we see that final examination marks in secondary education show a stable or slightly rising trend in the fields of Dutch language and mathematics, see figure 2.4. This finding contrasts with the Dutch results in PISA reading and mathematics, see section 2. However, there are a number of aspects that hinder a fair comparison, two of which are discussed in this text box.

Figure 2.4 Average final exam marks in secondary education by level of education for Dutch language (left) and mathematics (right)



Source: DUO (2014, 2020).

The curriculum that is tested in the national tests differs greatly from the international assessments (Swart et al., 2021; Van der Molen et al., 2019). For example, PISA reading focuses on the use of texts in social settings and for achieving personal goals, while the final exam programs mainly focus on abstract text analysis and text reflections. In addition, the final exams are not the same for each school type, while PISA makes no distinction between school types.

Both the final test in primary education and the final exams in secondary education are very important for a student's school career, while other (inter)national tests have no consequences. Students are more motivated for high-stakes tests. This is visible during the test (for example, whether the test is completed in full) but also in preparation for the test (for example, through test training). The final test of the eight grade in primary education showed, for example, that in 2018/2019 about 98% of students achieved level 1F in reading and about 93% 1F in mathematics (Inspectorate of Education, 2020). However, in the first grade of secondary education of 2018/2019, a quarter of the children failed level 1F in reading on a low-stakes test, and 45% were unable to achieve level 1F in mathematics (Inspectorate of Education, 2021). Although these are not exactly the same cohorts, the large difference suggests that motivation and preparation play a role. See section 3 for more information about the role of motivation.

3 Comparability across countries and over time is a challenge

Organizations that coordinate the international student assessments are faced with various factors that can undermine the reliability and comparability of the results. In this section we will discuss eight – what we call – equivalence problems: problems that impede a fair comparison of the scores between countries or between reference years, so that differences in scores do not always equate to differences in the knowledge and skills of pupils. An important disclaimer for this section is that there may be publication bias in the literature. Research that is critical of international student assessments may be published more easily and receive more attention than research that has not been able to detect any flaws. Moreover, the current literature is mainly based on the somewhat older reference years and it is not always clear to what extent the deficiencies have been resolved in later reference years.

Differences in difficulty level

Because a new test version is taken every reference year, the level of difficulty can vary from year to year. When the test versions from different reference years differ in terms of difficulty level, this complicates the comparability of the absolute scores over time. In order to compare the results from different reference years on the same scale, international assessments use anchor questions. These are questions that recur in several reference years and are used to correct for differences in difficulty. See, for example, the technical reports of the IEA (2020a, p. 15.1) and the OECD (2019b, p. 193). In addition, so-called link errors are published. Link errors reflect the uncertainty in the scores that arises from, among other things, the use of different test versions in different reference years (OECD, 2019b, p. 193-194).

Research shows that the selection of anchor questions can influence the outcomes. For example, Monseur and Berezner (2007) show that the average score of the OECD countries on the PISA 2003 reading test could have been four points higher or four points lower if a certain anchor question had been deleted. The most extreme case is Mexico, where deleting a certain anchor question could have resulted in a 16-point lower or 9-point higher score. This sensitivity did not affect Mexico's ranking: in both cases, Mexico would finish at the bottom of the ranking. However, this sensitivity had significant consequences for Japan. In the PISA 2003 reading test, Japan was ranked twelfth, but this could have been tenth or nineteenth if a particular anchor question had been dropped.

Differences in assessment framework

International student assessements do not test the same skills every reference year, or not to the same extent. Over the years, for example, what 'good reading' means has changed. In PIRLS, the emphasis has increasingly shifted to how well students are able to apply the information from texts to new situations, rather than demonstrating basic understanding (IEA, 2015, p. 11). And in PISA 2018, more attention was paid to reflection and evaluation in the reading component than in 2015: 30% versus 25% of the questions (OECD, 2019b, p.35). The OECD therefore indicates that score differences between reference years can also be the result of a changed assessment framework, and therefore not only arise because students have started to perform better or worse (OECD, 2019b, p.193).

Differences in language and culture

For linguistic and cultural reasons, it is difficult to produce fully comparable test versions for different countries. This can lead to differences in the degree of difficulty, which puts pressure on the comparability of

outcomes between countries. For example, Arffman (2010) argues that the translation of a source version into a target language is complicated by several problems, such as differences in word meaning, grammar and culture. Huang et al. (2016) give an appealing example of a test question that is bound to the social context. PISA 2006 included a test question that discussed a maglev train. This question turned out to be easier for students from China, where one maglev train is operational in 2006 and one under construction, than for students from Hong Kong, where such trains are not available.

The coordinating organizations try to reduce linguistic and cultural influences to a minimum by applying a careful translation procedure. The OECD uses a double translation and reconciliation procedure for the translation process: two persons independently translate the English and French source versions into the target language, after which a third person combines the two translations into one version (OECD, 2019c, p. 2). In addition, national project managers are allowed to let the translated text deviate from the source version in order to better reflect the local culture (OECD, 2019c, p. 8).

Several researchers conclude that the coordinating organizations fail to completely eliminate linguistic and cultural influences. An example is the study by Kreiner and Christensen (2014), who found that the 2009 PISA reading test is subject to differential item functioning (DIF) and that it is not possible to find a subset of questions that can withstand this. DIF means that a test question is answered better or worse by groups of comparable level, and is therefore an indicator of an equivalence problem. The authors come to the conclusion that the use of an adapted statistical method leads to different rankings. Studies focusing on PIRLS and TIMSS also question the comparability of scores between countries (Grisay et al., 2009; Karakoc Alatli et al., 2016).

Differences in motivation and endurance

Motivation and test endurance are partly culturally determined and can differ per reference year, which makes it difficult to compare skills. Not all students try equally hard and not all are equally good at keeping their attention on a test. However, a lack of motivation is not the same as a lack of knowledge or skills (Akyol et al., 2021). The fact that motivation and test endurance can differ between countries has been shown, inter alia, by a study by Borghans and Schils (2012) based on PISA. Pupils from Finland, Belgium and Austria, for example, have relatively high test endurance, while pupils from Greece, Italy and the United Kingdom have less endurance. Another factor is that PISA, TIMSS and PIRLS are low-stakes tests. This means that an individual student is not rewarded or judged based on the test result. Low-stakes tests can generally count on less motivation, although this can also differ per country (OECD, 2019b, p. 200). In addition, it is possible that motivation has decreased over the years, for example due to the increased intensity of tests (Penk et al., 2014).

Research shows that motivation and test endurance have a major impact on rankings. Research on motivation and test endurance usually focuses on PISA, and not on the tests taken by younger students. Borgonovi and Biecek (2016), for example, look at test endurance. In Finland, Estonia and Taipei, students have the highest test endurance. Estonia in particular therefore scores higher in the ranking. If only questions in the front of the booklet were looked at, Estonia would be 22nd in PISA Reading 2009. If only questions in the third cluster were looked at, they would be 13th. Their real ranking is 14th. Recent research by Akyol et al. (2021) looked at the influence of non-serious students on the score of PISA science 2015. This test was digital, so the researchers knew the time spent per question. According to the researchers, non-serious students click through questions faster than average. If all students in all countries had been serious, Portugal would have risen the most in the ranking, namely five places. According to Akyol et al., this indicates that there are many non-serious students in Portugal. The largest decline would be for the US, also with five places.

Coordinating organizations are aware of the role of motivation and test endurance, but only limited corrections are made. A large number of unanswered questions at the end of the test may indicate that a

student has lost motivation along the way. Unanswered questions at the end of the TIMSS test do not count towards the calculation of the score (IEA, 2020b, p. 10.5). In contrast, skipped questions in the middle of the TIMSS test are not ignored by default (IEA, 2020b, p. 10.5). Also in PISA, unanswered questions at the end of the test are not counted. Akyol et al. (2021) show that this partly removes the influence of motivation in PISA. Coordinating organizations also ask questions about a student's test motivation, for example: 'how much effort did you put into this test?'. Since the introduction of digital testing, it is also known how much time a student spends on a question (OECD, 2019b, pp. 200-201). However, this information is only used in additional analyses.

Differences in the composition of the student population

The main results of the international assessments do not take into account that the composition of the student population differs between reference years and between countries, but PISA does publish trend corrections for this in the appendix. Changes in the student population, for example due to migration, can contribute to differences in outcomes that are not related to the effectiveness of the education system (OECD, 2019b, p. 193). To take this into account, PISA publishes corrected trends in the appendix (OECD, 2019b, p. 140-141). These adjusted trends show that changes in the composition of the student population differ among countries. For example, a large part of the falling score on the reading test in Germany, Luxembourg, Norway and Switzerland can be explained by the changed composition. In addition, PISA publishes the aforementioned link errors (OECD, 2019b, p. 193-194), which reflect uncertainty due to the fact that the tests, among other things, are not performed by the same population every year. TIMSS and PIRLS take less account of changes in the student population. They publish trends in the average number of years that a student has been in education, the average age, the percentage of the target group that did not take the test and the percentage of participation after replacement (IEA, 2017, p. 5.26; IEA, 2020b, pp. 9.40). Because these characteristics have not changed much over the years for most countries, no corrected trends in scores are published.

Differences in the assessment mode

In recent years, the coordinating organizations have partly switched to computer-based testing, which can complicate a clear comparison over time and between countries. The PISA test has been digitized in steps. In 2015, the test was offered digitally for the first time, but in the form of plain text (OECD, 2016, p. 147). In 2018, the PISA test contained more digital elements, such as hyperlinks and tabs (OECD, 2019b, p. 182). The test was also administered adaptively, so that during the test the student is asked more difficult or easier questions, depending on the performance on previous questions (OECD, 2019b, p. 37). The TIMSS test was offered digitally for the first time in 2019 in a limited number of countries, including the Netherlands (IEA, 2020c, p. 1-2). The PIRLS test was still offered on paper in 2016, but countries could choose to take a separate computer-based test in addition to the paper-based test (IEA, 2017). The Netherlands did not participate in this. Computer-based testing offers a number of advantages, such as recording the time a student needs to answer a question and the possibility to make the test adaptive. At the same time, countries may differ in the level of digitization and the extent to which students are used to taking computer-based tests, which can affect test performance.

Prior to the switch to computer-based testing, the coordinating organizations carried out pilot tests, but the switch nevertheless seems to have had an impact on the scores. The pilot tests at TIMSS showed some differences between the paper- and computer-based versions. That is why some of the students still took the TIMSS 2019 test on paper (IEA, 2020b, p. 13.1). The OECD has also conducted pilot tests in which students were randomly given a paper- or computer-based version (OECD, 2017, p. 35) and in which students were randomly presented with the questions in a fixed order or adaptively (OECD, 2019d, h.2 p. 4-7). According to the OECD, the pilot tests did not show any major differences between the versions and, based on the outcomes, anchor questions were selected that scored the same in the different versions (OECD, 2017, pp. 152-162). However,

Jerrim et al. (2018) conclude that although the correction method in PISA has mitigated the problems, it has not completely eliminated them. An important underlying problem seems to be that the anchor questions linking 2015 to 2012 were selected based on the average outcome of the pilot test across all participating countries. However, at the country level some anchor questions scored differently in the computer-based test than in the paper version. To our knowledge, there is no research that focuses specifically on the transition to adaptive testing.

Differences in statistical methods

The OECD and IEA regularly update their methods to calculate the scores, which has an impact on the reported scores. Statistical techniques change and improve, and organizations move with the times. Model developments improve the accuracy of the scores in a reference year, but make it more difficult to compare between reference years (and possibly between countries). For example, in 2015, PISA moved to a new hybrid model that combines multiple statistical methods. This combined model was not used in previous years (OECD, 2017, p. 171). The methodological change worked out differently for different countries. For Estonia, for example, the reading score increased by 18 points between 2009 and 2015, while the increase would have been only 10 points if the 2015 method had been used in 2009 (OECD, 2016, p. 308). The IEA has also recently adjusted their statistical models to take into account changes in the assessment mode (IEA, 2020b, p. 11.14). In principle, the coordinating organizations do not correct for changes in statistical methods. For example, scores from previous reference years are not adjusted retroactively (OECD, 2019b, p. 193). To map out the impact, PISA does publish the aforementioned link errors (OECD, 2019b, p. 194).

Differences in representativeness of the sample

Non-response from schools and students may lead to results being based on a group of students that is not representative of the total population. Participation in international assessments by schools or students is voluntary in most countries. This can distort the results, for example if weaker schools or students more often refuse to participate. If the non-response differs per country or per reference year, this makes a clear comparison difficult. In order to overcome non-response as much as possible, when a school drops out, a replacement school with the same characteristics as the initial school is contacted. Furthermore, the coordinating organizations set target percentages for participation (OECD, 2019d, h.4, IEA, 2017, p. 3.1, IEA, 2020b, p. 3.1). If a country does not achieve this, the results are displayed at TIMSS and PIRLS with a disclaimer that they are not reliable and an additional analysis is done at PISA. Finally, the statistical models of PISA correct somewhat for student dropout by weighting the other participating students from the same school more heavily (OECD, 2019d, h.8). An important assumption here is that the students who dropped out are comparable to students from the same school who did participate.

International research shows that non-response affects the scores on the PISA test. For example, a recent study by Jerrim (2021) shows that adjusting for non-response would lower the UK's results in PISA 2018 by 10 to 15 points. Several researchers indicate that dropout of individual students is an important cause of the non-response bias in PISA (Micklewright et al., 2012; Monseur, 2005; Wuttke, 2007). The assumption that students who drop out are comparable to participating students from the same school turns out to be incorrect in many cases. For example, Jerrim (2021) shows that in the United Kingdom, students who perform less well, relatively often do not participate. However, because the OECD has hardly any background data on the students who did not participate, it cannot properly correct for this dropout.

Several researchers indicate that the coordinating organizations do not take into account uncertainty arising from non-response when calculating the standard errors. The reported standard errors are used, among other things, to determine whether the scores of countries differ significantly from each other and whether the score of a country has changed significantly over time. Rutkowski and Rutkowski (2016) and Schnepf (2018) indicate that the margins of error around the achieved scores are probably much larger than

the reported standard errors suggest. This means that differences over time and between countries are less often significant than stated in the reports.

4 Influence of equivalence problems on the score and ranking of the Netherlands

Because the literature is highly fragmented and often incomplete, it is not possible to capture the influence of all equivalence problems on the score and ranking of the Netherlands in one figure. The ranking of the Netherlands can also be influenced by equivalence problems that are not directly relevant to the Netherlands but do have an effect on the scores of competing countries. Nevertheless, a number of studies have been carried out that give us a rough picture of how the equivalence problems can influence the Dutch position in international student assessments. In this section we will discuss the main results of that literature. Where possible, we offer suggestions on how the equivalence problems can be presented more transparently and how they can be researched more extensively.

The literature provides indications that changes in the assessment mode, the assessment framework and the statistical methods used have had a negative impact on the position of the Netherlands. Feskens et al. (2019) investigated the change in the PISA assessment mode between 2012 and 2015, namely from a paper- to computer-based test. Based on an alternative scaling method, the authors conclude that Dutch mathematics results fell by 5 points between 2012 and 2015, which is less than half of the 12-point drop reported by PISA. However, Hamhuis et al. (2020) conclude that Dutch students at TIMSS found the computer-based test not more difficult than the paper version. With regard to changes in the assessment framework, the OECD indicates that these changes can lead to differences in scores between reference years (OECD, 2019b, p.193). This is relevant for the Netherlands, because PISA reading placed more emphasis on reflection and evaluation in 2018, which at the time coincided with a deterioration of the Dutch performance on this component. This degraded performance was therefore given greater weight by the shifting emphasis of the test. Adjusted statistical methods have also had an impact on the Dutch results. For example, it appears that the Dutch reading score in PISA would have increased by four points between 2009 and 2015 if the 2015 scaling method had been used in 2009 (OECD, 2016, p. 308). This contrasts with the five-point drop reported in official statistics, which equates to a nine-point difference.

International differences in test endurance and motivation seem to have positively influenced the position of the Netherlands. For example, Borgonovi and Biecek (2016) show that Dutch students have above average test endurance at PISA reading in 2009. If the score had been calculated on the basis of only the first questions, the Netherlands would have finished three places lower. A study by Borghans and Schils (2012) concludes that the test endurance of Dutch students on the PISA tests of 2003 and 2006 is fairly average. A recent study by Akyol et al. (2021) looked at the influence of non-serious students on the PISA science score in 2015. According to this study, there are relatively few non-serious students in the Netherlands, which has resulted in a higher ranking for the Netherlands. If all students in all countries were to be equally serious, the Netherlands would drop four places in the ranking. This is one of the largest declines in the study by Akyol et al.

Some studies indicate that the position of the Netherlands is sensitive to difficulty level and the way in which this is corrected for, although the effect is not unequivocally positive or negative. Monseur and Berezner (2007) show that the Dutch position has historically been influenced by the selection of anchor questions, which are used to correct for differences in difficulty. For example, the authors show that the score of the Netherlands on the PISA 2003 reading test (officially 513 points) could have been seven points higher or seven points lower if a certain anchor question had been deleted. The Netherlands would then have ended up two places higher or one place lower in the ranking (officially eighth place). A recent study by Cito (2021) examined whether the anchor questions in the PISA reference years 2012, 2015 and 2018 are sensitive to *differential item functioning* (DIF, see section 2) and what it would mean for the Dutch results if a correction is made for this. In the domain of reading, the correction leads to a significantly higher score in 2018. However, the higher score is not enough to offset the decrease between 2015 and 2018. In the domain of mathematics, the correction does not lead to significantly different outcomes. For science, the correction leads to a significantly lower score in the reference year 2015.

The position of the Netherlands may be sensitive to non-representativeness, but the effect is unclear and more research is needed for definitive conclusions. Non-representativeness is an important point of concern for the Netherlands because the Netherlands regularly fails to meet the target percentages for school participation. This happened, for example, in 2015 and 2018 at PISA and in 2019 at TIMSS. Nevertheless, the results from these reference years have been included in the final report. Additional analyzes showed that the realized sample did not differ from the total population in terms of visible characteristics (Anders et al., 2021; Meelissen et al., 2020). In addition, dropout at the student level is also a point of concern. The Netherlands does meet the target percentages of the OECD on this point, but nevertheless performs relatively weak on this aspect. Given the findings in the international literature (see section 3), the high dropout rate at the student level could be problematic for the reliability of the Dutch score. Hard conclusions cannot be drawn, however, because research into the consequences of non-response at the student level is not available in the Netherlands. This requires linking the PISA data to administrative education data, which is not yet possible. We will come back to this later in this section.

A striking development that is specifically relevant to the Netherlands is the sharp increase in the use of one-hour booklets. In the reference years up to and including 2015, approximately 2% to 3% of Dutch students received a one-hour booklet for the PISA test. By 2018, however, this had increased to as much as 18 percent. The reason for this increase was a pilot test that showed that many students in remedial education (in Dutch: *leerwegondersteunend onderwijs*) struggled with the regular PISA test. To accommodate these students, it was decided in 2018 to give these students a shortened and simpler version of the test. In addition to the sharp increase in the use of these modified booklets, the OECD also changed the statistical method used to determine the scores of the Dutch one-hour booklets in 2018. PISA provides a brief analysis of this change and found no evidence that the change affected the score of the Netherlands (OECD, 2019d, annex H). To our knowledge, the increased use of the one-hour booklets and the modified statistical method has not been investigated by other researchers.

There is no evidence that the absolute score of the Netherlands is substantially influenced by the composition of the student population or differences in language and culture. The influence of a changing socio-economic composition of the student population seems to play a limited role in the Netherlands. According to the corrected scores published by PISA, the performance of the Netherlands is only zero to two points lower when corrected for the composition of the student population (OECD, 2019b, p. 254). However, these effects are significant for a number of other countries, so that it could have indirectly influenced the ranking of the Netherlands (OECD, 2019b, p. 141). We are not aware of any literature that provides insight into the consequences of language and cultural differences for the position of the Netherlands.

A striking finding is that the sharp decline in the Dutch score in PISA reading between 2015 and 2018 coincided with a significant number of equivalence problems. For example, during that period PISA took further steps towards a computer-based assessment mode. Also, adaptive testing was introduced. Research by Feskens et al. (2019) has shown that similar changes between 2012 and 2015 influenced the score of the Netherlands. In addition, the assessment framework for the reading domain has been adjusted – with more attention to critical evaluation, on which the Netherlands performs relatively less well – and research by Cito (2021) showed that a correction for DIF in the reference year 2018 would have led to a higher reading score. Finally, in 2018 we saw a remarkably large increase in the use of one-hour booklets and the target percentage for participation by schools was not achieved by the Netherlands. It is unclear what the net effect of all these factors was on the position of the Netherlands. We can say, however, that the position of the Netherlands is subject to larger-than-usual uncertainty.

Linking the results of international student assessments to other educational outcomes can provide important insights into the importance and nature of equivalence problems for the Netherlands. At the moment it is not possible to link the data from PISA, TIMSS and PIRLS to other educational outcomes of the same students, such as the final test in primary education, the student tracking system and the final exam in secondary education. Linking these data can provide insight into the representativeness of the sample, which is an important point for concern, especially in the Netherlands because of the high non-response and the use of one-hour booklets. A comparison of the results of the international assessments with the results of the final test or final exams can also provide insight into the extent to which motivation plays a role.

Although equivalence problems are in part unavoidable, there is still something to be gained in improving the transparency of the problems and the choices that are made. It would be good if the coordinating organizations provide more insight into the consequences of important choices that are inevitably somewhat arbitrary. This includes, for example, the selection of anchor questions – which are used to correct for differences in the difficulty level and the switch to computer-based and adaptive testing – and assumptions in the statistical modelling. In addition, the coordinating organizations can do more to provide insight into the influence of changes in the assessment framework. For example, by also publishing results that correct for a shifting emphasis of the test. This is already done in other areas: in PISA 2018, the figures in the appendix are corrected for changes in the student population and statistical methods.

References

Akyol, S.P., K. Krishna and J. Wang, 2021, Taking PISA Seriously: How Accurate Are Low Stakes Exams?, *Journal of Labor Research*, vol. 42, pag. 184-243.

Anders, J., S. Has, J. Jerrim, N. Shure and L. Zieger, 2021, Is Canada really an education superpower? The impact of non-participation on results from PISA 2015, *Educational Assessment, Evaluation and Accountability*, vol. 33, nr. 1, pag. 229-249.

Araujo, L., A. Saltelli and S.V. Schnepf, 2017, Do PISA data justify PISA-based education policy?, *International Journal of Comparative Education and Development*, vol. 19, nr. 1, pag. 20-34.

Arffman, I., 2010, Equivalence of translations in international reading literacy studies, *Scandinavian Journal of Educational Research*, vol. 54, nr. 1, pag. 37-59.

Baird, J., T. Isaacs, S. Johnson, G. Stobart, G. Yu, T. Sprague and R. Daugherty, 2011, *Policy effects of PISA*, Bristol: University of Bristol.

Borghans, L. and T. Schils, 2012, *The leaning tower of Pisa: decomposing achievement test scores into cognitive and noncognitive components*, Working Paper, Maastricht: Maastricht University.

Borgonovi, F. and P. Biecek, 2016, An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test, *Learning and Individual Differences*, vol. 49, pag. 128-137.

Breakspear, S., 2014, How does PISA shape education policy making? Why how we measure learning determines what counts in education, *Centre for Strategic Education Seminar Series*, vol. 240.

Cito, 2021, *PISA Trends within the Netherlands – Evaluating mode effects*, Arnhem: Stichting Cito.

DUO, 2014, *Examenmonitor VO 2013*, Groningen: Dienst Uitvoering Onderwijs.

DUO, 2020, *Examenmonitor VO 2019*, Groningen: Dienst Uitvoering Onderwijs.

Feskens, R., J.P. Fox and R. Zwitser, 2019, Differential item functioning in PISA due to mode effects. In: Veldkamp, B.P. en C. Sluijter (eds.), *Theoretical and practical advances in computer-based educational measurement* (pag. 232-247), Cham: Springer.

Grisay, A., E. Gonzalez and C. Monseur, 2009, Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessements, *IERI*, vol. 2, pag. 63-83.

Hamhuis, E., C. Glas and M. Meelissen, 2020, Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students?, *British journal of educational technology*, vol. 51, nr. 6, pag. 2340-2358.

Huang, X., M. Wilson and L. Wang, 2016, Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture, *Educational Psychology*, vol. 36, nr. 2, pag. 378-390.

IEA, 2009, *PIRLS 2011 Assessment Framework*, Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

IEA, 2015, *PIRLS 2016 Assessment Framework*, Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

IEA, 2017, *Methods and Procedures in PIRLS 2016*, Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

IEA, 2020a, *Developing the TIMSS 2019 Mathematics and Science Achievement Instruments*, Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

IEA, 2020b, *Methods and Procedures: TIMSS 2019 Technical Report*, Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

IEA, 2020c, *TIMSS 2019: International Results in Mathematics and Science*, Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Inspectorate of Education, 2020, *De Staat van het Onderwijs 2020*, Utrecht: Inspectorate of Education.

Inspectorate of Education, 2021, *De Staat van het Onderwijs 2021*, Utrecht: Inspectorate of Education.

Jerrim, J., J. Micklewright, J.H. Heine, C. Salzer and C. McKeown, 2018, PISA 2015: how big is the 'mode effect' and what has been done about it?, *Oxford Review of Education*, vol. 44, nr. 4, pag. 476-493.

Jerrim, J., 2021, PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK?, *Review of Education*, vol. 9, nr. 3, pag. E3270.

Karakoc Alatli, B., C. Ayan, B. Polat Demir and G. Uzun, 2016, Examination of the TIMSS 2011 Fourth Grade Mathematics Test in Terms of Cross-Cultural Measurement Invariance, *Eurasian Journal of Educational Research*, vol. 66, pag. 389-406.

Kreiner, S. and K.B. Christensen, 2014, Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy, *Psychometrika*, vol. 79, nr. 2, pag. 210-231.

Meelissen, M., E. Hamhuis and L. Weijn, 2020, *Leerlingprestaties in de exacte vakken in groep 6 van het basisonderwijs: Resultaten TIMSS-2019*, Twente: University of Twente.

Micklewright, J., S.V. Schnepf and C. Skinner, 2012, Non-response biases in surveys of schoolchildren: the case of the English Programme for International Student Assessment (PISA) samples, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 175, nr. 4, pag. 915-938.

Molen, P. van der, S. Schouwstra, R. Feskens and M. van Onna, 2019, *Vaardigheidsontwikkelingen volgens PISA en examens*, Arnhem: Stichting Cito Instituut voor Toetsontwikkeling.

Monseur, C., 2005, An exploratory alternative approach for student non response weight adjustment, *Studies in Educational Evaluation*, vol. 31, nr. 2-3, pag. 129-144.

Monseur, C. and A. Berezner, 2007, The computation of equating errors in international surveys in education, *Journal of Applied Measurement*, vol. 8, nr. 3, pag. 323-335.

OCW, 2022, *Kamerbrief over Masterplan basisvaardigheden*, The Hague: Ministry of Education, Culture and Science.

OECD, 2016, *PISA 2015 Results (Volume I): Excellence and Equity in Education*, Paris: OECD Publishing.

OECD, 2017, *PISA 2015 Technical report*, Paris: OECD Publishing.

OECD, 2019a, *PISA 2018 Assessment and Analytical Framework*, Paris: OECD Publishing.

OECD, 2019b, *PISA 2018 Results (Volume I): What Students Know and Can Do*, Paris: OECD Publishing.

OECD, 2019c, *PISA 2021. Translation and Adaptation Guidelines*, Vienna: First Meeting of the PISA 2021 National Project Managers.

OECD, 2019d, *PISA 2018 Technical Report*, Paris: OECD Publishing.

Dutch Education Council, 2022, *Taal en rekenen in het vizier*, The Hague: Dutch Education Council.

Penk, C., C. Pöhlmann and A. Roppelt, 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*, vol. 2, nr. 5.

Rutkowski, L. and D. Rutkowski, 2016, A call for a more measured approach to reporting and interpreting PISA results, *Educational Researcher*, vol. 45, nr. 4, pag. 252-257.

Schnepf, S., 2018, Insights into survey errors of large scale educational achievement surveys, *JRC Working Papers in Economics and Finance (No. 2018/5)*.

Swart, N., I. van Barneveld, A. van der Lee, C. Dood and J. Gubbels, 2021, *Aanvullende analyses op PISA-2018 data: Antwoord op onderzoeksvraag 5*, Nijmegen: Expertisecentrum Nederlands.

Wiseman, A.W., 2013, Policy responses to PISA in comparative perspective. In: Meyer, H-D & A. Benavot (eds.), *PISA, Power, and Policy: the emergence of global educational governance*, pag. 303-322, Southampton: Hobbs the Printers.

Wuttke, J., 2007, Uncertainty and Bias in PISA, In: Hopmann, S.T., G. Brinek en M. Retzl (eds.), *PISA according to PISA: Does PISA keep what it promises*, pag. 241-263, Münster: LIT Verlag.