



Een blik op de Nederlandse positie in internationale onderwijsrankings

Internationaal peilingsonderzoek duidt op een verslechtering van de basisvaardigheden onder Nederlandse jongeren.

In deze publicatie onderzoeken we in hoeverre deze onderzoeken een goede graadmeter zijn voor onderwijsprestaties.

We constateren dat de positie van Nederland gevoelig is voor instroom van nieuwe deelnemende landen en methodologische problemen.

Het is onbekend of de optelsom van deze problemen de positie van Nederland gunstig of ongunstig beïnvloedt.

Om internationaal peilingsonderzoek beter te duiden en bruikbaarder te maken, zou het behulpzaam zijn als de uitvoerende organisaties de impact van methodologische keuzes uitgebreider analyseren en het mogelijk maken hun resultaten te koppelen aan nationale toetsen.

CPB - december 2022

Paul Verstraten,
Annikka Lemmens,
Marielle Non

Samenvatting

Volgens internationaal peilingsonderzoek daalt het niveau van de basisvaardigheden in Nederland.

Hierover zijn de afgelopen jaren zorgen ontstaan. In het driejaarlijkse onderzoek *Programme for International Student Assessment* (PISA) onder 15-jarige middelbare scholieren zien we sinds 2015 een verslechtering van de score van Nederland op het gebied van lezen en natuurwetenschappen. De score voor PISA wiskunde daalt al sinds 2003, maar vertoonde in 2018 (het laatste peiljaar) wel een licht herstel. Ook ten opzichte van andere landen verslechtert de positie van Nederland bij PISA. Onderzoek van *Trends in International Mathematics and Science Study* (TIMSS), dat zich richt op 10-jarigen, toont sinds 1995 een licht dalende score van Nederland bij natuurwetenschappen en wiskunde. De positie van Nederland op deze TIMSS-ranking neemt ten opzichte van andere landen sterk af. Het onderzoek *Progress in International Reading Literacy Study* (PIRLS) laat daarentegen zien dat de leesvaardigheden van 10-jarigen tussen 2001 en 2016 redelijk stabiel zijn gebleven. Mede door deze peilingsonderzoeken is er momenteel aandacht voor het niveau van de basisvaardigheden in Nederland.

Uit dit onderzoek blijkt dat internationaal peilingsonderzoek niet altijd een accurate graadmeter is voor onderwijsprestaties.

Dit komt doordat verschillende problemen de score en positie van een land kunnen beïnvloeden. Zo blijkt bijvoorbeeld dat veranderingen in de afnamemodus, zoals de overstap van papieren naar digitale toetsing, voor sommige landen gunstiger uitpakken dan voor andere. Dit maakt de vergelijking van de scores over de tijd en tussen landen lastiger. De uitvoerende organisaties achter de peilingsonderzoeken maken wel gebruik van statistische methoden om te corrigeren voor verschillen in de afnamemodus, maar slagen hier niet altijd volledig in. Bovendien voeren deze organisaties regelmatig veranderingen en verbeteringen door in de statistische methoden. Dergelijke aanpassingen komen de betrouwbaarheid van de cijfers in de nieuwe peiljaren ten goede, maar maken een vergelijking met eerdere peiljaren lastiger. Verder constateren we dat de uitvoerende organisaties regelmatig het toetsingskader herzien. Zo is door de jaren heen veranderd wat ‘goed kunnen lezen’ inhoudt. Dit heeft voordelen, maar ook dit bemoeilijkt het vergelijken van uitkomsten met die van eerdere jaren. Ook verschillen tussen landen in de motivatie waarmee leerlingen de toets maken en gebreken bij de steekproef kunnen de betrouwbaarheid en vergelijkbaarheid beïnvloeden. De optelsom van deze problemen staat een zuivere vergelijking van de scores tussen landen en tussen peiljaren in de weg. Om deze reden staan verschillen in scores niet altijd gelijk aan verschillen in kennis en vaardigheden van leerlingen.

De literatuur laat zien dat deze problemen de positie van Nederland zowel positief als negatief kunnen beïnvloeden. Het blijkt niet mogelijk om het totaaleffect van de problemen in een cijfer te vangen. Een aantal studies laat zien dat veranderingen in de afnamemodus, het toetsingskader en de statistische methoden de positie van Nederland bij PISA wiskunde in 2015 en PISA lezen in 2015 en 2018 ongunstig hebben beïnvloed. Ander onderzoek heeft gekeken naar de invloed van motivatie op de score van Nederland. Hieruit blijkt dat Nederlandse leerlingen gemiddeld serieuzer zijn dan leerlingen in andere landen, wat de score van Nederland bij PISA natuurwetenschappen in 2015 positief heeft beïnvloed. Verschillen in motivatie staan echter niet gelijk aan verschillen in kennis of kunde. Tot slot zijn er aanwijzingen dat non-respons mogelijk invloed heeft op de positie van Nederland. Het effect hiervan op de Nederlandse positie is echter onduidelijk. Doordat de literatuur sterk versplinterd en vaak onvolledig is, is het niet mogelijk om de invloed van al deze factoren op de score en ranking van Nederland in een cijfer te vangen.

De internationale peilingsonderzoeken zouden een aantal wijzigingen kunnen doorvoeren die de hierboven genoemde problematiek transparanter maken en die bijdragen aan oplossingen. Ten eerste zou het nuttig zijn als de data van PISA kunnen worden gekoppeld aan nationale toetsen (zoals de eindtoets en het eindexamen) en het leerlingvolgsysteem. Dit is momenteel niet mogelijk, wat bijvoorbeeld onderzoek naar de kwaliteit van de steekproef belemmert. Ten tweede zou het goed zijn als de peilingsonderzoeken

transpanter zijn over de gevolgen van belangrijke keuzes die onvermijdelijk enigszins arbitrair zijn. Hierbij valt te denken aan aannames bij de statistische modellering. Ten derde zouden de internationale peilingsonderzoeken meer kunnen doen om de invloed van wijzigingen in het toetsingskader inzichtelijk te maken.

1 Inleiding

Internationaal peilingsonderzoek monitort de cognitieve vaardigheden van leerlingen uit verschillende landen. De meest uitvoerige internationale peilingsonderzoeken zijn het *Programme for International Student Assessment* (PISA), *Trends in International Mathematics and Science Study* (TIMSS) en *Progress in International Reading Literacy Study* (PIRLS). PISA wordt driejaarlijks uitgevoerd door de *Organisatie voor Economische Samenwerking en Ontwikkeling* (OESO) en TIMSS en PIRLS worden respectievelijk vier- en vijfjaarlijks uitgebracht door de *International Association for the Evaluation of Educational Achievement* (IEA). Deze onderzoeken richten zich op verschillende vakgebieden en leeftijdscategorieën. PIRLS focust op leesvaardigheden onder 10-jarigen en TIMSS richt zich op rekenen en natuurwetenschappen bij dezelfde leeftijd.¹ PISA bestrijkt de domeinen lezen, wiskunde en natuurwetenschappen onder 15-jarigen.

Landen grijpen de uitkomsten van internationaal peilingsonderzoek veelvuldig aan om grote beleidsmatige wijzigingen in het onderwijs door te voeren. Een forse beleidsaanpassing als reactie op onverwacht slechte PISA-resultaten wordt in de literatuur het ‘PISA-shockverschijnsel’ genoemd (Breakspear, 2014). Baird e.a. (2011) noemen o.a. Frankrijk als voorbeeld, waar als reactie op de resultaten van PISA 2009 werd aangekondigd dat het basisonderwijs zich meer zou gaan focussen op basisvaardigheden en gepersonaliseerd onderwijs. Een ander voorbeeld waaruit de invloed van internationaal peilingsonderzoek blijkt, is de casus van Noorwegen. Toen in 2000 en 2003 bleek dat Noorwegen beneden het PISA-gemiddelde presteerde, voerde het land hervormingen door op het gebied van het curriculum en de toetsing. Volgens Wiseman (2013) hebben ook Duitsland en Japan in het verleden grote beleidswijzigingen doorgevoerd vanwege tegenvallende resultaten bij internationaal peilingsonderzoek.

De afgelopen jaren zijn ook in Nederland zorgen ontstaan over de verslechterende positie van Nederland in internationale peilingsonderzoeken. De meest recente scores van Nederland, van eind jaren '10, zijn namelijk bij alle internationale peilingsonderzoeken lager dan de score die Nederland noteerde in het eerste peiljaar, medio jaren '90, en begin jaren '00. Naast de daling in de score, heeft Nederland ook te maken met een verslechterende positie ten opzichte van andere landen. Waar Nederland twee decennia geleden nog kon bogen op een solide top 10-positie bij alle internationale peilingsonderzoeken, is dit nu alleen nog het geval bij PISA wiskunde en natuurwetenschappen. Zorgen over deze ontwikkeling werden geuit in onder andere De Staat van het Onderwijs (Inspectie van het Onderwijs, 2021, p. 98) en de kamerbrief over het Masterplan basisvaardigheden (OCW, 2022, p. 3). Critici merken overigens op dat deze peilingsonderzoeken voorbijgaan aan het belang van niet-cognitieve vaardigheden, zoals relationele vaardigheden en doorzettingsvermogen. De resultaten van internationaal peilingsonderzoek zouden daarom slechts een beperkte graadmeter zijn voor onderwijskwaliteit (Araujo e.a., 2017).

Op verzoek van het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) heeft het CPB onderzocht in hoeverre internationaal peilingsonderzoek een accurate graadmeter is voor de Nederlandse onderwijsprestaties. Onze analyse bestaat uit drie onderdelen. In paragraaf 2 analyseren we de ontwikkeling van Nederland bij internationale peilingsonderzoeken. In paragraaf 3 bieden we een overzicht van equivalentieproblemen bij internationaal peilingsonderzoek die het lastig maken om landen onderling en over de tijd te vergelijken. Tot slot gaan we in paragraaf 4 in op de consequenties van deze equivalentieproblemen voor de positie van Nederland.

¹ We laten TIMSS voor voortgezet onderwijs buiten beschouwing omdat Nederland sinds 2008 niet meer deelneemt aan dit onderzoek.

2 De positie van Nederland in internationaal peilingsonderzoek

In deze paragraaf kijken we naar de positie van Nederland in internationaal peilingsonderzoek aan de hand van drie maatstaven: de absolute score, de totale ranking en een consistente ranking. De absolute score betreft de puntenscore van Nederland. De totale ranking toont de positie van Nederland ten opzichte van alle andere deelnemende landen. De toename van deelnemende landen over tijd beïnvloedt deze totale ranking. Daarom tonen we ook een consistente ranking: een ranking met enkel landen die in dezelfde peiljaren als Nederland hebben deelgenomen.² Deze ranking wordt niet vertekend door instroom en uitstroom van deelnemers, maar heeft als nadeel dat bij sommige rankings maar weinig landen overblijven.

Lezen

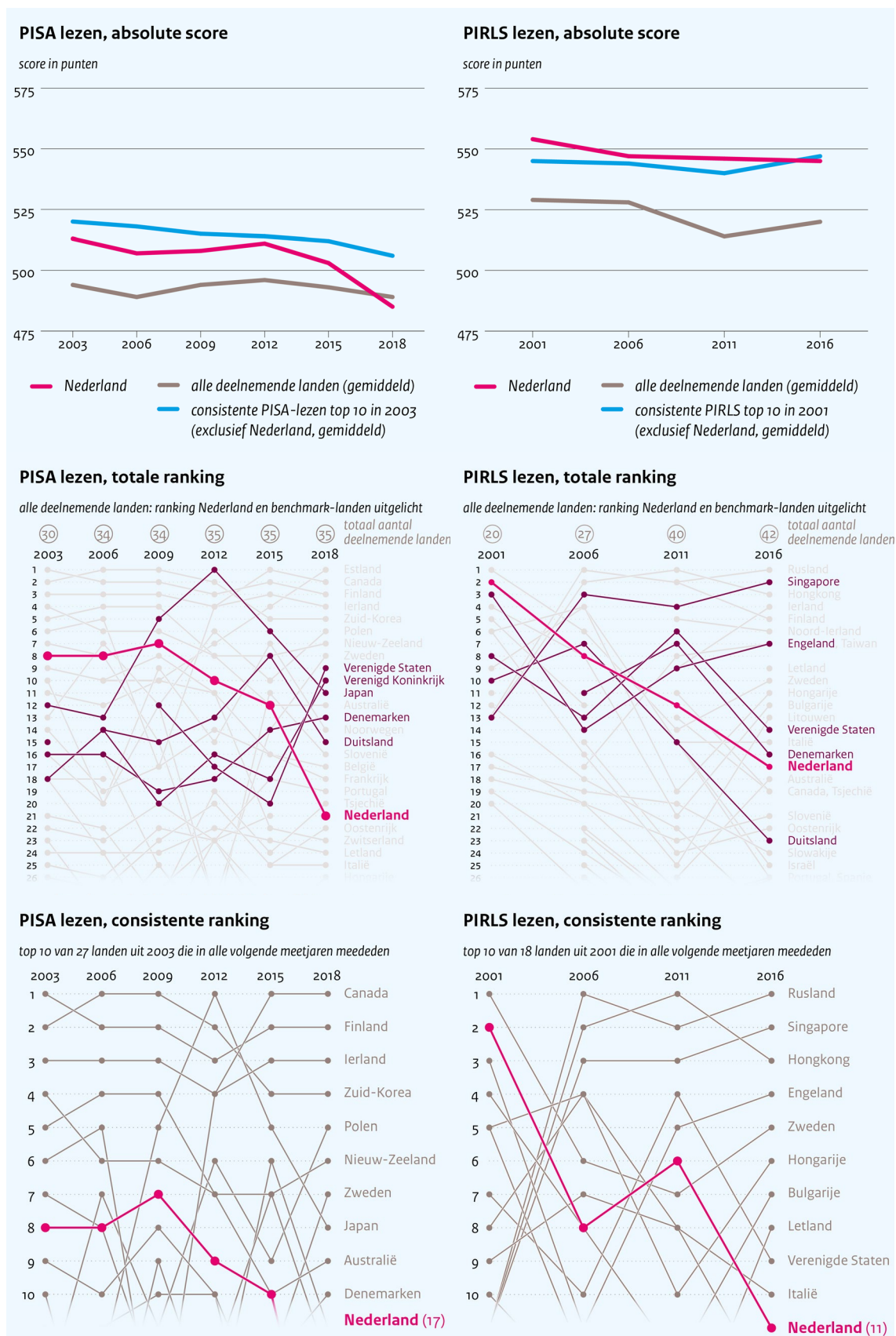
De absolute score van Nederland bij PISA lezen was tot 2012 vrij stabiel, maar is daarna aanzienlijk gedaald, waardoor Nederland op zowel de totale als de consistente ranking een duikeling heeft gemaakt. De linkerhelft van figuur 2.1 toont de ontwikkeling van Nederland bij PISA lezen. De score van Nederland bleef tot 2012 op peil, maar zette daarna een daling in die tussen 2015 en 2018 zelfs 18 punten bedroeg. Nederland scoorde daarmee in 2018 rond het gemiddelde van alle deelnemers. De top 10 best presterende landen laat over tijd een lichte daling zien in de absolute score. De score van Nederland is in de laatste twee peiljaren echter veel sterker gedaald, waardoor Nederland aanzienlijk is afgeleden op de ranking. In de ranking zijn een aantal benchmark-landen uitgelicht die qua economische ontwikkeling redelijk vergelijkbaar zijn met Nederland. Ook ten opzichte van deze landen is Nederland fors gedaald. De instroom van nieuwe landen verklaart slecht in zeer beperkte mate de daling van Nederland in de ranking. Ook de consistente ranking laat immers een forse daling zien, van plaats 8 in 2003 naar plaats 17 in 2018.

Nederlandse 10-jarigen scoren op de leestoets van PIRLS – na een lichte daling tussen 2001 en 2006 – relatief stabiel, maar op de ranking zakt Nederland wel steeds verder af doordat andere landen hun score verbeteren en er meer landen deelnemen. De rechterhelft van figuur 2.1 toont de ontwikkeling van Nederland bij PIRLS. Nederland scoort in de eerste peiljaren vrij hoog in vergelijking met de best presterende landen. Tussen 2011 en 2016 hebben andere landen hun score echter verbeterd, terwijl Nederlandse leerlingen stabiel zijn blijven presteren. Hierdoor is Nederland op de consistente ranking een aantal plaatsen gezakt. Op de totale ranking is Nederland sterker gezakt, wat naast de stijging in de score van andere landen ook te verklaren is door een toename van het aantal deelnemers.

De door PISA gemeten daling in de leesvaardigheid van 15-jarigen tussen 2012 en 2018 contrasteert scherp met de vrijwel constante score van 10-jarigen bij PIRLS tussen 2006 en 2016. Deze toetsen hebben grotendeels betrekking op dezelfde cohorten leerlingen. Een mogelijke verklaring voor het verschil kan zijn dat de leesvaardigheden zich tussen het tiende en vijftiende levensjaar relatief zwak hebben ontwikkeld of dat de motivatie sterk is teruggelopen. Een andere verklaring kan zijn dat de onderzoeken niet hetzelfde meten. Zo geeft PISA 2018 meer gewicht aan het onderdeel evalueren en reflecteren dan PIRLS 2011 en 2016 (IEA, 2009, p. 14; IEA, 2015, p. 14; OESO, 2019a, p. 42). Paragraaf 3 gaat dieper in op de verschillende aspecten die het beeld van internationaal peilingsonderzoek kunnen vertekenen.

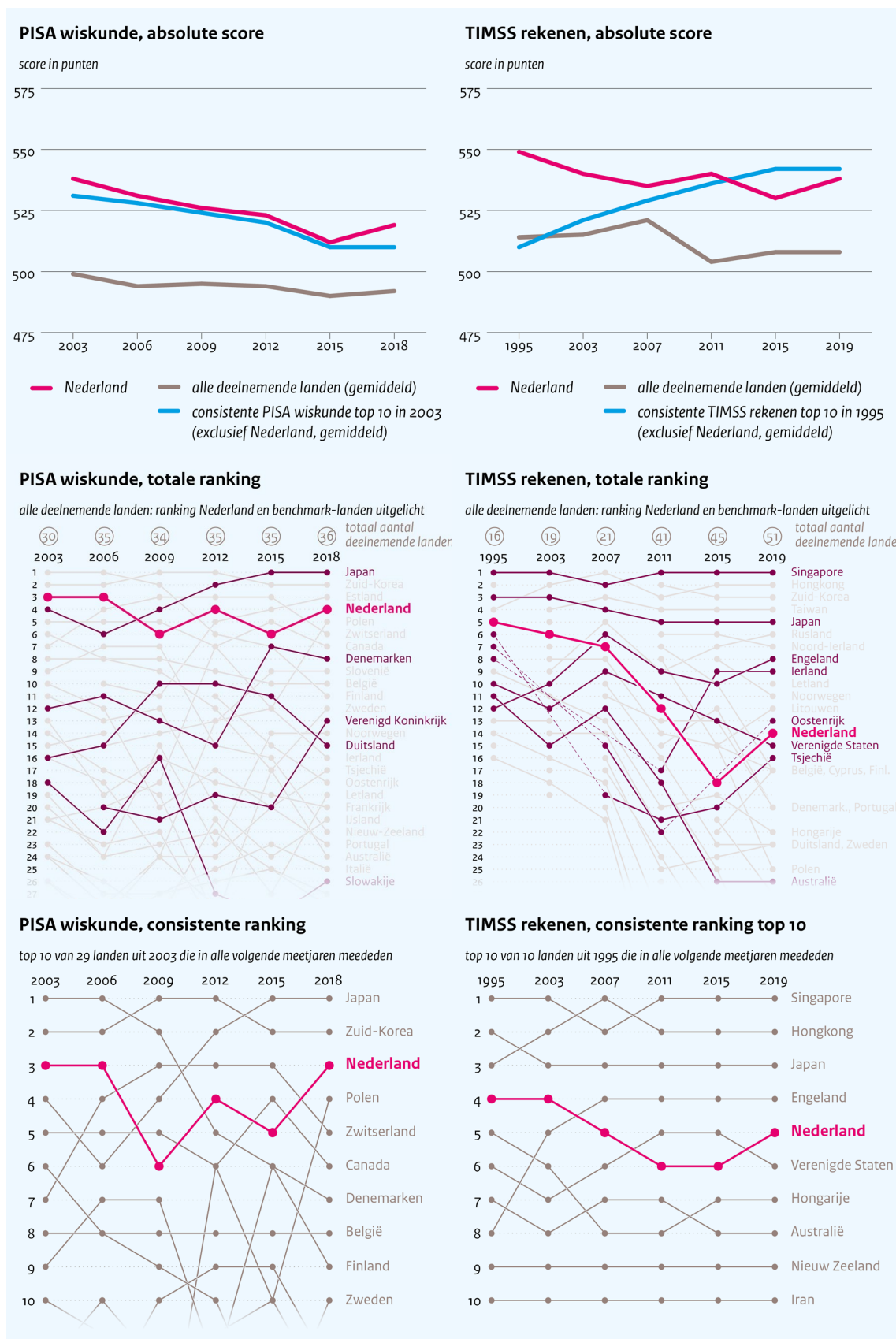
² Nederland heeft vrijwel alle peiljaren van PISA, TIMSS en PIRLS meegedaan. Nederland heeft ook deelgenomen aan PISA 2000, maar de non-respons was te hoog, waardoor er geen resultaten voor Nederland zijn gepubliceerd. In 1999 heeft Nederland bij TIMSS alleen meegedaan bij de toets voor voortgezet onderwijs. Deze resultaten laten we buiten beschouwing.

Figuur 2.1 Leesvaardigheden bij PISA (15-jarigen, links) en PIRLS (10-jarigen, rechts)



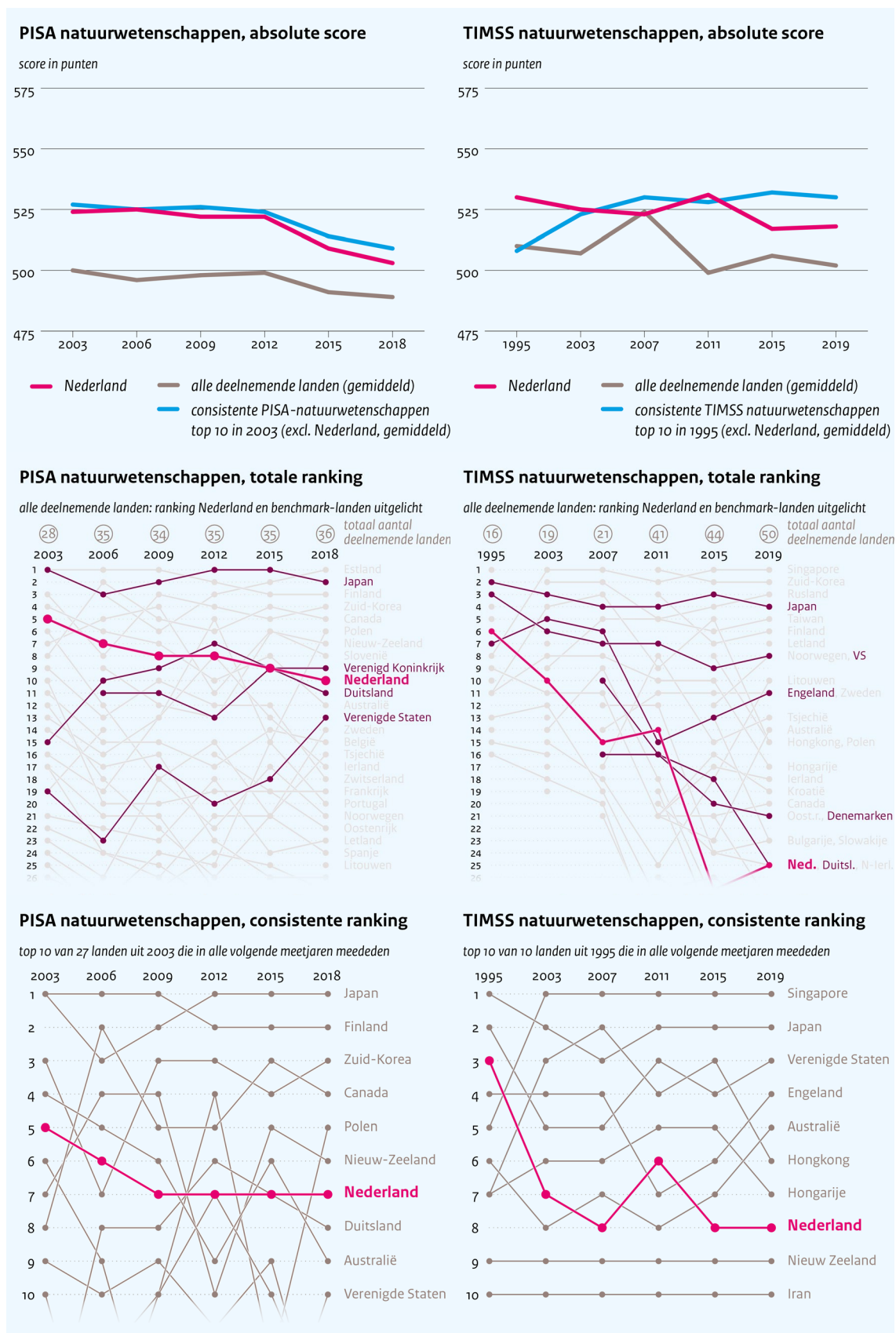
Bron: eigen analyses op basis van onderwijsincijfers.nl (PISA, links) en het Download Center van de IEA (PIRLS, rechts).

Figuur 2.2 Wiskunde en rekenen bij PISA (15-jarigen, links) en TIMSS (10-jarigen, rechts)



Bron: eigen analyses op basis van onderwijscijfers.nl (PISA, links) en het Download Center van de IEA (TIMSS, rechts).

Figuur 2.3 Natuurwetenschappen bij PISA (15-jarigen, links) en TIMSS (10-jarigen, rechts)



Bron: eigen analyses op basis van onderwijscijfers.nl (PISA, links) en het Download Center van de IEA (TIMSS, rechts).

Wiskunde/rekenen en natuurwetenschappen

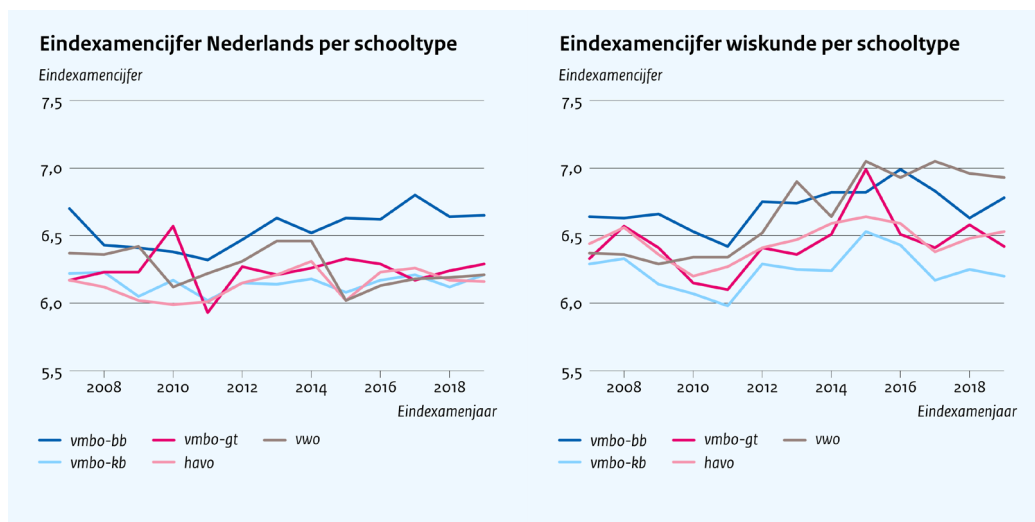
Bij PISA wiskunde en natuurwetenschappen vertoont Nederland een dalende trend in de absolute score, maar blijft de top 10-klassering behouden. De linkerhelft van figuur 2.2 en 2.3 toont de ontwikkeling van Nederland bij respectievelijk PISA wiskunde en natuurwetenschappen. De absolute score bij wiskunde vertoont al vanaf het eerste peiljaar een dalende trend, hoewel de score in 2018 weer iets is gestegen. Bij natuurwetenschappen is de score met name de laatste twee peiljaren sterk gedaald. De top 10 goed presterende landen laat echter een vergelijkbare dalende trend zien, waardoor Nederland in zowel de totale als de consistente ranking slechts beperkt daalt. De verschillen tussen de toptanden zijn klein, waardoor de ranking een onregelmatig patroon heeft. Nederland scoorde bijvoorbeeld in 2012 bij natuurwetenschappen een punt hoger dan Australië en twee punten lager dan Duitsland. In beide gevallen was het verschil insignificant. In 2015 scoorde Nederland juist een punt lager dan Australië en gelijk aan Duitsland. Een kleine verandering in absolute score kan dus al leiden tot een stijging of daling van een of meerdere plaatsen in de ranking.

De absolute score van Nederland bij TIMSS rekenen en natuurwetenschappen is licht dalend, maar Nederland daalt aanzienlijk op de ranking doordat de top 10-landen beter zijn gaan presteren en er over tijd een forse instroom van deelnemers is geweest. De rechterhelft van figuur 2.2 en 2.3 toont de ontwikkeling van Nederland bij respectievelijk TIMSS rekenen en natuurwetenschappen. De uitkomsten in 1995 en 2019 zijn met meer dan de gebruikelijke onzekerheid omgeven omdat Nederland in die peiljaren niet het vereiste aantal deelnemende scholen heeft bereikt. De absolute score fluctueert enigszins, maar laat over de hele periode genomen een lichte daling zien. Andere goed presterende landen laten een dergelijke daling niet zien. In combinatie met een stijging in het aantal deelnemende landen, leidt dit ertoe dat Nederland sterk is gezakt op de totale ranking: bij rekenen van plaats 5 in 1995 naar plaats 14 in 2019 en bij natuurwetenschappen van plaats 6 in 1995 naar plaats 25 in 2019. De consistente ranking laat een meer bescheiden daling zien, maar dit beeld kan vertekend zijn omdat er maar weinig landen zijn die vanaf het eerste peiljaar altijd hebben meegedaan.

Nationale cijfers tonen een ander beeld, maar zijn lastig te vergelijken

Nationale cijfers – van centrale eindexamens, eindtoetsen en nationale peilingen – schetsen deels een ander beeld van de basisvaardigheden in Nederland. Nationale cijfers over de beheersing van taal en rekenen zijn schaars in Nederland (Onderwijsraad, 2022), maar de cijfers die wel beschikbaar zijn geven niet altijd hetzelfde beeld als internationaal peilingsonderzoek. Zo zien we bijvoorbeeld dat eindexamencijfers een stabiele of licht stijgende trend op het gebied van Nederlands en wiskunde vertonen (zie figuur 2.4). Deze bevinding staat in contrast met de Nederlandse resultaten bij PISA lezen en wiskunde (zie paragraaf 2). Er zijn echter een aantal aspecten die een zuivere vergelijking belemmeren, waarvan we er twee in dit tekstkader bespreken.

Figuur 2.4 Gemiddelde eindexamencijfers per onderwijsniveau voor Nederlands (links) en wiskunde (rechts)



Bron: DUO (2014, 2020).

Het curriculum dat de nationale toetsen testen verschilt sterk van dat in de internationale toetsen (Swart e.a., 2021; Van der Molen e.a., 2019). Zo focust PISA lezen op het gebruik van teksten in het maatschappelijke verkeer en voor het behalen van persoonlijke doelen, terwijl de eindexamenprogramma's vooral gaan over abstracte tekstanalyse en tekstbeschouwingen. Voor de eindexamens komt daar nog bij dat die verschillen per schooltype, terwijl PISA geen onderscheid maakt tussen schooltypes.

Zowel de eindtoets als de eindexamens zijn zeer belangrijk voor de schoolloopbaan van een leerling, terwijl andere (inter)nationale toetsen hier geen gevolgen voor hebben. Leerlingen doen beter hun best voor *high stakes*-toetsen dan voor toetsen waar weinig van afhangt. Dit is zichtbaar tijdens de toets (of de toets bijvoorbeeld volledig wordt ingevuld) maar ook in de voorbereiding op de toets (bijvoorbeeld via de inzet van toetstraining). Uit de eindtoets van groep acht bleek bijvoorbeeld dat in 2018/2019 ongeveer 98% van de leerlingen niveau 1F voor lezen behaalde en ongeveer 94% 1F voor rekenen (Inspectie van het Onderwijs, 2020). In de brugklas van 2018/2019 bleek echter een kwart van de kinderen niveau 1F voor lezen niet te behalen bij een *low-stakes*-toets, en 45% was niet in staat niveau 1F voor rekenen te behalen (Inspectie van het Onderwijs, 2021). Dit zijn weliswaar niet precies dezelfde cohorten, maar het grote verschil doet vermoeden dat motivatie en voorbereiding een rol spelen. Zie paragraaf 3 voor meer informatie over de rol van motivatie.

3 Vergelijkbaarheid tussen landen en over tijd is een uitdaging

Uitvoerende organisaties achter de internationale peilingsonderzoeken hebben te maken met diverse factoren die de betrouwbaarheid en vergelijkbaarheid van de uitkomsten kunnen ondermijnen. In deze paragraaf bespreken we een achttal – wat wij noemen – equivalentieproblemen. Dit zijn problemen die een zuivere vergelijking van de scores tussen landen of tussen peiljaren belemmeren, waardoor verschillen in scores niet altijd gelijkstaan aan verschillen in kennis en vaardigheden van leerlingen. Een belangrijke disclaimer bij deze paragraaf is dat er mogelijk sprake is van publicatiebias in de literatuur. Onderzoek dat kritisch is op internationaal peilingsonderzoek wordt namelijk mogelijk sneller gepubliceerd en krijgt meer aandacht dan onderzoek dat geen gebreken heeft kunnen ontdekken. Bovendien is de huidige literatuur voornamelijk gebaseerd op de wat oudere peiljaren en is het niet altijd duidelijk in hoeverre de onvolkomenheden zijn opgelost in latere peiljaren.

Verschillen in de moeilijkheidsgraad

Omdat er ieder peiljaar een nieuwe toetsversie wordt afgenomen, kan de moeilijkheidsgraad van jaar tot jaar verschillen. Wanneer de toetsversies uit verschillende peiljaren qua moeilijkheidsgraad van elkaar verschillen, vormt dit een complicerende factor voor de vergelijkbaarheid van de absolute scores over de tijd. Om de uitkomsten uit verschillende peiljaren op eenzelfde schaal te vergelijken, maken internationale peilingsonderzoeken daarom gebruik van ankervragen. Dit zijn vragen die in meerdere peiljaren terugkomen, en die worden gebruikt om te corrigeren voor verschillen in de moeilijkheidsgraad. Zie bijvoorbeeld de verantwoording van de IEA (2020a, p. 15.1) en de OESO (2019b, p. 193). Daarnaast publiceren de uitvoerende organisaties zogeheten *link errors*. *Link errors* geven de onzekerheid in de scores weer die ontstaat door onder andere het gebruik van verschillende toetsversies in verschillende peiljaren (OESO, 2019b, p. 193-194).

Onderzoek toont aan dat de selectie van ankervragen invloed kan hebben op de uitkomsten. Zo laten Monseur en Berezner (2007) zien dat de gemiddelde score van de OESO-landen op de PISA 2003-leestoets zowel vier punten hoger als vier punten lager had kunnen uitvallen als een bepaalde ankervraag zou zijn geschrapt. Het meest extreme geval is Mexico, waarbij het schrappen van een bepaalde ankervraag had kunnen leiden tot een zestien punten lagere of negen punten hogere score. Deze gevoeligheid had echter geen invloed op de ranking van Mexico: in beide gevallen zou Mexico onderaan de ranglijst eindigen. Deze gevoeligheid had echter wel aanzienlijke gevolgen voor Japan. Bij de leestoets van PISA 2003 behaalde Japan een twaalfde plek, maar dit had ook de tiende of negentiende plek kunnen zijn als een bepaalde ankervraag was geschrapt.

Verschillen in het toetsingskader

Internationale peilingsonderzoeken toetsen niet ieder peiljaar dezelfde vaardigheden, of niet in dezelfde mate. Zo is door de jaren heen veranderd wat ‘goed kunnen lezen’ inhoudt. Bij PIRLS is de nadruk steeds meer gaan liggen op hoe goed leerlingen in staat zijn om de informatie uit teksten toe te passen op nieuwe situaties, in plaats van op het laten zien van basisbegrip (IEA, 2015, p. 11). En bij PISA 2018 is bij het onderdeel lezen meer aandacht geschonken aan reflecteren en evalueren dan in 2015: 30% versus 25% van de vragen ging hierover (OESO, 2019b, p. 35). De OESO geeft daarom aan dat scoreverschillen tussen peiljaren ook het gevolg kunnen zijn van een veranderd toetsingskader, en dus niet alleen ontstaan doordat leerlingen beter of slechter zijn gaan presteren (OESO, 2019b, p. 193).

Verschillen in taal en cultuur

Om taalkundige en culturele redenen is het lastig om voor verschillende landen volledig vergelijkbare toetsversies te produceren. Dit kan leiden tot verschillen in de moeilijkheidsgraad, die de vergelijkbaarheid van de uitkomsten tussen landen onder druk zetten. Zo stelt Arffman (2010) dat het vanwege meerdere factoren moeilijk is om een bronversie van een toets naar een doeltaal te vertalen. Verschillen in woordbetekenis, grammatica en cultuur bemoeilijken dit proces bijvoorbeeld. Huang e.a. (2016) geven een aansprekend voorbeeld van een toetsvraag die gebonden is aan de maatschappelijke context. In PISA 2006 was een toetsvraag opgenomen waarin werd gesproken over een magneetzweeftrein. Deze vraag bleek eenvoudiger voor leerlingen uit China, waar in 2006 één magneetzweeftrein operationeel was en één in aanbouw, dan voor leerlingen uit Hong Kong, waar dergelijke treinen niet beschikbaar zijn.

De uitvoerende organisaties proberen taalkundige en culturele invloeden tot een minimum te beperken door de toepassing van een zorgvuldige vertaalprocedure. De OESO gebruikt voor het vertaalproces een dubbele vertalings- en afstemmingsprocedure. Dit houdt in dat twee personen onafhankelijk van elkaar de Engelstalige en Franstalige bronversies naar de doeltaal vertalen, waarna een derde persoon de twee vertalingen verenigt in één versie (OESO, 2019c, p. 2). Daarnaast is het nationale projectmanagers toegestaan om de vertaalde tekst van de bronversie te laten afwijken om daarmee beter aan te sluiten op de lokale cultuur (OESO, 2019c, p. 8).

Meerdere onderzoekers komen tot de conclusie dat de uitvoerende organisaties er niet in slagen om taalkundige en culturele invloeden volledig uit te schakelen. Een voorbeeld is de studie van Kreiner en Christensen (2014), die vaststelt dat de PISA leestoets uit 2009 onderhevig is aan *differential item functioning* (DIF) en dat het niet mogelijk is een subset van vragen te vinden die hiertegen bestand is. DIF betekent dat een toetsvraag beter of slechter wordt beantwoord door groepen die qua niveau vergelijkbaar zijn. Daarmee is DIF een indicator voor een equivalentieprobleem. De auteurs komen tot de conclusie dat het gebruik van een aangepaste statistische methode leidt tot andere rankings. Ook studies die zich richten op PIRLS en TIMSS plaatsen vraagtekens bij de vergelijkbaarheid van scores tussen landen (Grisay e.a., 2009; Karakoc Alatlı e.a., 2016).

Verschillen in motivatie en uithoudingsvermogen

Motivatatie en testuithoudingsvermogen zijn deels cultureel bepaald en kunnen verschillen per peiljaar, wat het vergelijken van vaardigheden lastig maakt. Leerlingen doen niet allemaal even hard hun best en kunnen niet allemaal even goed hun aandacht bij een toets houden. Gebrek aan motivatie is echter niet hetzelfde als een gebrek aan kennis of kunde (Akyol e.a., 2021). Dat motivatie en testuithoudingsvermogen kunnen verschillen tussen landen blijkt onder andere uit een onderzoek van Borghans en Schils (2012) op basis van PISA. Leerlingen uit bijvoorbeeld Finland, België en Oostenrijk hebben een relatief hoog testuithoudingsvermogen, terwijl leerlingen uit Griekenland, Italië, en het Verenigd Koninkrijk minder lang hun best doen. Verder speelt het ook een rol dat PISA, TIMSS en PIRLS *low-stakes*-toetsen zijn. Dit betekent dat een individuele leerling niet wordt beloond of afgerekend op basis van de testuitslag. Voor *low-stakes*-toetsen hebben leerlingen over het algemeen minder motivatie, al kan ook dit per land verschillen (OESO, 2019b, p. 200). Daarnaast is het mogelijk dat de motivatie is afgenomen over de jaren, bijvoorbeeld door de toegenomen intensiteit van toetsen (Penk e.a., 2014).

Onderzoek wijst uit dat motivatie en testuithoudingsvermogen een grote impact hebben op de rankings. Onderzoek naar motivatie en testuithoudingsvermogen richt zich meestal op PISA, en niet op de toetsen voor jongere leerlingen. Borgonovi en Biecek (2016) kijken bijvoorbeeld naar testuithoudingsvermogen. In Finland, Estland en Taipei hebben leerlingen het hoogste testuithoudingsvermogen. Vooral Estland scoort daardoor hoger op de ranking. Als alleen naar vragen voor in het boekje wordt gekeken, dan zou Estland 22^{ste} zijn bij PISA lezen 2009. Zou er alleen naar vragen in het derde cluster worden gekeken, dan zijn ze 13^{de}. Hun echte

ranking is 14^{de}. In recent onderzoek van Akyol e.a. (2021) is gekeken naar de invloed van niet-serieuze leerlingen op de score bij natuurwetenschappen van PISA 2015. Deze toets was digitaal, waardoor de onderzoekers de bestede tijd per vraag kennen. Niet-serieuze leerlingen klikken volgens de onderzoekers bovengemiddeld snel door vragen heen. Als alle leerlingen in alle landen serieus zouden zijn geweest, dan zou Portugal het sterkst stijgen op de ranking, namelijk vijf plaatsen. Dit geeft volgens Akyol e.a. aan dat er in Portugal veel niet-serieuze leerlingen zijn. De grootste daler zou de VS zijn, ook met vijf plaatsen.

Uitvoerende organisaties zijn zich bewust van de rol van motivatie en testuithoudingsvermogen, maar er vinden slechts beperkt correcties plaats. Een groot aantal onbeantwoorde vragen aan het einde van de toets kan een aanwijzing zijn dat een leerling gaandeweg de toets de motivatie verloren is. 'Niet-bereikte' vragen aan het einde van de TIMSS-toets tellen niet mee bij het berekenen van de score (IEA, 2020b, p. 10.5). Overgeslagen vragen in het midden van de TIMSS-toets blijven daarentegen niet standaard buiten beschouwing (IEA, 2020b, p. 10.5). Ook bij PISA tellen onbeantwoorde vragen aan het eind van de toets niet mee. Akyol e.a. (2021) laten zien dat dit de invloed van motivatie bij PISA deels wegneemt. Uitvoerende organisaties stellen daarnaast vragen over de toetsmotivatie van een leerling. Zo vragen ze bijvoorbeeld: 'Hoeveel moeite heb je in deze toets gestoken?'. Ook is sinds de introductie van digitale toetsing bekend hoeveel tijd een leerling aan een vraag besteedt (OESO, 2019b, p. 200-201). Deze informatie gebruiken de organisaties echter alleen voor aanvullende analyses.

Verschillen in de samenstelling van de leerlingpopulatie

De hoofdresultaten van de peilingsonderzoeken houden er geen rekening mee dat de samenstelling van de leerlingpopulatie verschilt tussen peiljaren en tussen landen, maar PISA publiceert hiervoor wel trendcorrecties in de appendix. Wijzigingen in de populatie, bijvoorbeeld door migratie, kunnen bijdragen aan verschillen in uitkomsten die niet samenhangen met de effectiviteit van het onderwijssysteem (OESO, 2019b, p. 193). Om hier rekening mee te houden, publiceert PISA in de appendix gecorrigeerde trends (OESO, 2019b, p. 140-141). Uit deze gecorrigeerde trends blijkt dat veranderingen in de samenstelling van de leerlingpopulatie verschillend uitpakt voor de diverse landen. Zo is een groot deel van de dalende score op de leestoets in Duitsland, Luxemburg, Noorwegen en Zwitserland te verklaren door de veranderde samenstelling. Daarnaast publiceert PISA de eerdergenoemde *link errors* (OESO, 2019b, p. 193-194). Deze geven onder andere de onzekerheid in de scores weer die ontstaat doordat de toetsen niet ieder jaar door dezelfde populatie worden gemaakt. TIMSS en PIRLS houden minder rekening met wijzigingen in de leerlingpopulatie. Zij publiceren trends in het gemiddeld aantal jaren dat een leerling onderwijs heeft gevolgd, de gemiddelde leeftijd, het percentage van de doelgroep dat niet meedeelde aan de toets en het percentage deelname na vervanging (IEA, 2017, p. 5.26; IEA, 2020b, p. 9.40). Omdat deze kenmerken door de jaren heen voor de meeste landen niet veel veranderd zijn, publiceren deze onderzoeken geen gecorrigeerde trends in scores.

Verschillen in de afnamemodus

De uitvoerende organisaties zijn de afgelopen jaren deels overgestapt op digitaal toetsen, wat een zuivere vergelijking over tijd en tussen landen kan bemoeilijken. De PISA-toets is in stappen gedigitaliseerd. In 2015 werd de toets voor het eerst digitaal aangeboden, in de vorm van platte tekst (OESO, 2016, p. 147). In 2018 bevatte de PISA-toets meer digitale elementen, zoals hyperlinks en tabbladen (OESO, 2019b, p. 182). Ook werd de toets toen adaptief afgenomen, wat inhoudt dat de leerling gedurende de toets moeilijkere of makkelijkere vragen krijgt afhankelijk van de prestaties op eerdere vragen (OESO, 2019b, p. 37). De TIMSS-toets is in 2019 voor het eerst digitaal aangeboden in een beperkt aantal landen, waaronder Nederland (IEA, 2020c, p. 1-2). De PIRLS-toets werd in 2016 nog op papier aangeboden, maar landen konden er wel voor kiezen om naast de papieren versie een separate digitale toets af te nemen (IEA, 2017). Nederland heeft hier niet aan meegedaan. Digitaal toetsen biedt een aantal voordelen, zoals registratie van de tijd die een leerling nodig heeft voor het beantwoorden van een vraag en de mogelijkheid om de toets adaptief te maken.

Tegelijk kunnen er verschillen bestaan tussen landen in het niveau van digitalisering en de mate waarin leerlingen gewend zijn om toetsen digitaal te maken. Dat kan de prestaties op de toets beïnvloeden.

Voorafgaand aan de overstap naar digitaal toetsen hebben de uitvoerende organisaties proeftesten gedaan, maar desondanks lijkt de overstap invloed te hebben gehad op de scores. De proeftesten bij TIMSS lieten enkele verschillen zien tussen de papieren en digitale versie. Daarom heeft een deel van de leerlingen de TIMSS 2019-toets nog op papier gemaakt (IEA, 2020b, p. 13.1). Ook de OESO heeft proeftesten gedaan waarbij leerlingen willekeurig een papieren of digitale versie kregen (OESO, 2017, p. 35) en waarbij leerlingen willekeurig de vragen in een vaste volgorde of adaptief kregen aangeboden (OESO, 2019d, h. 2 p. 4-7). Volgens de OESO lieten de proeftesten geen grote verschillen zien tussen de versies. Op basis van de uitkomsten hebben zij ankervragen geselecteerd die in de verschillende versies gelijk scoorden (OESO, 2017, p. 152-162). Jerrim e.a. (2018) concluderen echter dat de correctiemethode bij PISA de problemen wel heeft verzacht, maar niet volledig heeft weggenomen. Een belangrijk onderliggend probleem lijkt te zijn dat de ankervragen die 2015 aan 2012 linken zijn geselecteerd op basis van de gemiddelde uitkomst van de proeftest over alle deelnemende landen. Op landniveau scoren sommige ankervragen echter anders in de digitale toets dan in de papieren versie. Voor zover bij ons bekend is er geen onderzoek dat zich specifiek richt op de overstap naar adaptief toetsen.

Verschillen in statistische methoden

De OESO en IEA updaten regelmatig hun methoden om scores te berekenen, wat zijn weerslag heeft op de gerapporteerde scores. Statistische technieken veranderen en verbeteren, en de organisaties gaan met hun tijd mee. Modelontwikkelingen komen de nauwkeurigheid van de scores in een peiljaar ten goede, maar bemoeilijken de vergelijking tussen peiljaren (en mogelijk tussen landen). In 2015 is PISA bijvoorbeeld overgestapt op een nieuw hybride model dat meerdere statistische methoden combineert. In de voorgaande jaren maakte PISA geen gebruik van dit gecombineerde model (OESO, 2017, p. 171). De wijziging in methode pakte verschillend uit voor verschillende landen. Zo nam voor Estland de leesscore tussen 2009 en 2015 toe met 18 punten, terwijl de toename slechts 10 punten was geweest als in 2009 de methode van 2015 was gebruikt (OESO, 2016, p. 308). Ook de IEA heeft de statistische modellen recentelijk aangepast zodat ze rekening kunnen houden met veranderingen in de afnamemodus (IEA, 2020b, p. 11.14). In principe corrigeren de uitvoerende organisaties niet voor veranderingen in statistische methoden. Ze passen scores uit voorgaande jaren bijvoorbeeld niet retroactief aan (OESO, 2019b, p. 193). Om de impact in kaart te brengen publiceert PISA wel de eerdergenoemde *link errors* (OESO, 2019b, p. 194).

Verschillen in non-representativiteit

Non-respons van scholen en leerlingen kan ertoe leiden dat de resultaten zijn gebaseerd op een groep leerlingen die niet representatief is voor de totale populatie. Deelname aan internationale toetsen door scholen of leerlingen is in de meeste landen vrijwillig. Dit kan de resultaten vertekenen, bijvoorbeeld als zwakkere scholen of leerlingen vaker deelname weigeren. Als de non-respons verschilt per land of per peiljaar, maakt dit een zuivere vergelijking moeilijk. Om non-respons zo veel mogelijk te ondervangen, schrijven de uitvoerende organisaties bij uitval van een school een vervangende school aan met dezelfde kenmerken als de initiële school. Verder stellen de uitvoerende organisaties streefpercentages voor deelname (OESO, 2019d, h. 4, IEA, 2017, p. 3.1, IEA, 2020b, p. 3.1). Als een land deze niet behaalt, geven TIMSS en PIRLS de resultaten weer met een disclaimer dat deze niet betrouwbaar zijn en PISA voert een aanvullende analyse uit. Als laatste corrigeren de onderliggende statistische modellen bij PISA enigszins voor uitval van leerlingen door de andere deelnemende leerlingen van dezelfde school zwaarder mee te wegen (OESO, 2019d, h. 8). Een belangrijke aanname hierbij is dat de uitgevallen leerlingen vergelijkbaar zijn met leerlingen van dezelfde school die wel mee hebben gedaan.

Internationaal onderzoek toont aan dat non-respons invloed heeft op de scores op de PISA-toets. Een recente studie van Jerrim (2021) laat bijvoorbeeld zien dat correctie voor non-respons de resultaten van het Verenigd Koninkrijk bij PISA 2018 met 10 tot 15 punten zou laten dalen. Verschillende onderzoekers geven aan dat uitval van individuele leerlingen een belangrijke oorzaak is van de vertekening door non-respons bij PISA (Micklewright e.a., 2012; Monseur, 2005; Wuttke, 2007). De aanname dat uitgevallen leerlingen vergelijkbaar zijn met deelnemende leerlingen van dezelfde school blijkt in veel gevallen niet correct. Zo toont Jerrim (2021) aan dat slechter presterende leerlingen in het Verenigd Koninkrijk relatief vaak niet meedoen. Omdat de OESO echter nauwelijks achtergrondgegevens heeft van de leerlingen die niet hebben meegedaan, kan zij niet goed corrigeren voor deze uitval.

Verschillende onderzoekers geven aan dat de uitvoerende organisaties bij de berekening van de standaardfouten geen rekening houden met onzekerheid die voortkomt uit non-respons. De uitvoerende organisaties gebruiken de gerapporteerde standaardfouten onder meer om te bepalen of de scores van landen significant van elkaar verschillen en of de score van een land significant is veranderd over tijd. Rutkowski en Rutkowski (2016) en Schnepf (2018) geven aan dat de foutmarges rondom de behaalde scores naar alle waarschijnlijkheid een stuk groter zijn dan de gerapporteerde standaardfouten doen voorkomen. Dit betekent dat verschillen over tijd en tussen landen minder vaak significant zijn dan vermeld in de rapportages.

4 Invloed equivalentieproblemen op de score en ranking van Nederland

Doordat de literatuur sterk versplinterd en vaak onvolledig is, is het niet mogelijk om de invloed van alle equivalentieproblemen op de score en ranking van Nederland in een cijfer te vangen. De ranking van Nederland kan bovendien worden beïnvloed door equivalentieproblemen die niet direct relevant zijn voor Nederland maar wel een effect hebben op de score van concurrerende landen. Toch zijn er wel een aantal studies uitgevoerd die ons een grof beeld geven van hoe de equivalentieproblemen de Nederlandse positie bij internationaal peilingsonderzoek kunnen beïnvloeden. In deze paragraaf bespreken we op hoofdlijnen de belangrijkste uitkomsten van die literatuur. Waar mogelijk bieden we suggesties hoe de equivalentieproblemen transparanter kunnen worden gepresenteerd en beter kunnen worden onderzocht.

De literatuur biedt aanwijzingen dat veranderingen in de afnamemodus, het toetsingskader en de toegepaste statistische methoden een negatieve invloed hebben gehad op de positie van Nederland. Feskens e.a. (2019) onderzochten de wijziging in de PISA-afnamemodus tussen 2012 en 2015, namelijk van een papieren naar digitale toets. Op basis van een alternatieve schalingsmethode trekken de auteurs de conclusie dat de Nederlandse wiskunderesultaten tussen 2012 en 2015 zijn gedaald met 5 punten, wat ruim de helft minder is dan de door PISA gerapporteerde daling van 12 punten. Hamhuis e.a. (2020) concluderen echter dat Nederlandse studenten bij TIMSS niet meer moeite hadden met een digitale toetsafname dan met een papieren test. Met betrekking tot veranderingen in het toetsingskader geeft de OESO aan dat deze kunnen leiden tot scoreverschillen tussen peiljaren (OESO, 2019b, p. 193). Dit is relevant voor Nederland, omdat PISA lezen in 2018 meer nadruk legde op reflecteren en evalueren, wat toentertijd samenviel met een verslechtering van de Nederlandse prestaties op dit onderdeel. Deze verslechterde prestatie kreeg door de focusverschuiving dus een groter gewicht. Ook aangepaste statistische methoden hebben hun weerslag gehad op de Nederlandse resultaten. Zo blijkt dat de Nederlandse leesscore bij PISA tussen 2009 en 2015 met vier punten zou zijn toegenomen als in 2009 de schalingsmethode van 2015 was gebruikt (OESO, 2016, p. 308). Dit staat in contrast met de in officiële statistieken gerapporteerde daling van vijf punten, wat neerkomt op een verschil van negen punten.

Internationale verschillen in testuithoudingsvermogen en motivatie lijken de positie van Nederland juist positief te hebben beïnvloed. Zo laten Borgonovi en Biecek (2016) bijvoorbeeld zien dat Nederlandse leerlingen beschikken over een bovengemiddeld hoog testuithoudingsvermogen bij PISA lezen in 2009. Als de score zou zijn berekend op basis van alleen de eerste vragen, was Nederland drie plaatsen lager geëindigd. Een onderzoek van Borghans en Schils (2012) komt tot de conclusie dat het testuithoudingsvermogen van Nederlandse leerlingen op de PISA-toetsen van 2003 en 2006 redelijk gemiddeld is. In recent onderzoek van Akyol e.a. (2021) is gekeken naar de invloed van niet-serieuze leerlingen op de score bij natuurwetenschappen van PISA 2015. Volgens dit onderzoek zijn er in Nederland relatief weinig niet-serieuze leerlingen, wat ertoe heeft geleid dat Nederland hoger is geëindigd op de ranking. Als alle leerlingen in alle landen in dezelfde mate serieus zouden zijn, dan zou Nederland vier plaatsen dalen op de ranking. Dit is een van de grootste dalingen uit het onderzoek van Akyol e.a.

Sommige studies wijzen erop dat de positie van Nederland gevoelig is voor de moeilijkheidsgraad en de manier waarop daarvoor wordt gecorrigeerd, hoewel het effect niet eenduidig positief of negatief is.

Monseur en Berezner (2007) tonen aan dat de selectie van ankervragen de Nederlandse positie in het verleden heeft beïnvloed. De uitvoerende organisaties gebruiken ankervragen om te corrigeren voor verschillen in de moeilijkheidsgraad. Zo laten de auteurs zien dat de score van Nederland op de PISA 2003-leestoets (officieel 513 punten) zowel zeven punten hoger als zeven punten lager had kunnen uitvallen als een bepaalde ankervraag zou zijn geschrapt. Nederland zou dan op de ranking (officieel plaats acht) respectievelijk twee plaatsen hoger of een plaats lager zijn geëindigd. In recent onderzoek van Cito (2021) is gekeken of de ankervragen in de PISA-peiljaren 2012, 2015 en 2018 gevoelig zijn voor *differential item functioning* (DIF, zie paragraaf 2) en wat het zou betekenen voor de Nederlandse resultaten wanneer hiervoor wordt gecorrigeerd. Op het domein lezen leidt de correctie tot een significant hogere score in 2018. De hogere score is echter niet voldoende om de daling tussen 2015 en 2018 teniet te doen. Op het domein wiskunde leidt de correctie niet tot significant andere uitkomsten. Bij natuurwetenschappen leidt de correctie tot een significant lagere score in het peiljaar 2015.

De positie van Nederland is mogelijk gevoelig voor non-representativiteit, maar het effect is onduidelijk en voor definitieve conclusies is meer onderzoek nodig.

Non-representativiteit is voor Nederland een belangrijk punt van zorg omdat Nederland met enige regelmaat de streefpercentages voor deelname van scholen niet haalt. Dit gebeurde bijvoorbeeld in 2015 en 2018 bij PISA en in 2019 bij TIMSS. De resultaten uit deze peiljaren zijn wel in de eindrapportage opgenomen. Uit aanvullende analyses bleek namelijk dat de gerealiseerde steekproef niet verschilde van de totale populatie qua zichtbare kenmerken (Anders e.a., 2021; Meelissen e.a., 2020). Daarnaast is ook uitval op leerlingniveau een aandachtspunt. Nederland haalt op dit punt wel de streefpercentages van de OESO, maar behoort desalniettemin tot 'de slechtste jongetjes van de klas'. Gezien de bevindingen in de internationale literatuur (zie paragraaf 3), zou de hoge uitval op leerlingniveau problematisch kunnen zijn voor de betrouwbaarheid van de Nederlandse score. Harde conclusies kunnen op dit vlak echter niet worden getrokken, omdat onderzoek naar de gevolgen van non-respons op leerlingniveau in Nederland niet beschikbaar is. Hiervoor is een koppeling van de PISA-data aan administratieve gegevens nodig, wat vooralsnog niet mogelijk is. We komen hier later in deze paragraaf op terug.

Een opvallende ontwikkeling die specifiek voor Nederland relevant is, is het sterk toegenomen gebruik van de één-uurs-boekjes. In de peiljaren tot en met 2015 kreeg circa 2% tot 3% van de Nederlandse leerlingen bij de PISA-toets een één-uurs-boekje (*une heure booklet*). In 2018 was dit toegenomen tot wel 18%. De aanleiding van deze toename was een proeftest waaruit bleek dat veel leerlingen uit het leerwegondersteunend onderwijs moeite hadden met de reguliere PISA-toets. Om deze leerlingen tegemoet te komen, is er in 2018 voor gekozen om deze leerlingen een ingekorte en eenvoudigere versie van de toets te geven. Naast de sterke toename in het

gebruik van deze aangepaste boekjes, heeft de OESO in 2018 ook de statistische methode gewijzigd waarmee de scores van de Nederlandse één-uurs-boekjes worden bepaald. PISA biedt een beknopte analyse van deze wijziging en vond geen bewijs dat de wijziging invloed heeft gehad op de score van Nederland (OESO, 2019d, annex H). Voor zover bij ons bekend, is het toegenomen gebruik van de één-uurs-boekjes en de aangepaste statistische methode niet door andere partijen onderzocht.

Er is geen bewijs dat de samenstelling van de leerlingpopulatie of verschillen in taal en cultuur de Nederlandse score substantieel hebben beïnvloed. De invloed van een veranderende sociaaleconomische samenstelling van de leerlingpopulatie lijkt in Nederland een beperkte rol te spelen. Volgens de door PISA gepubliceerde gecorrigeerde scores vallen de prestaties van Nederland slechts nul tot twee puntscore lager uit wanneer je corrigeert voor de samenstelling van de leerlingpopulatie (OESO, 2019b, p. 254). Deze effecten zijn voor een aantal andere landen echter wel aanzienlijk, waardoor ze indirect alsnog de ranking van Nederland beïnvloed zouden kunnen hebben (OESO, 2019b, p. 141). Er is bij ons geen literatuur bekend die inzicht biedt in de gevolgen van taal- en cultuurverschillen voor de Nederlandse positie.

Een opvallende constatering is dat de scherpe daling van de Nederlandse score bij PISA lezen tussen 2015 en 2018 samenviel met een aanzienlijk aantal equivalentieproblemen. Zo zijn er bij PISA in die periode verdere stappen gezet richting een digitale afnamemodus en is ook het adaptief toetsen geïntroduceerd. Onderzoek van Feskens e.a. (2019) heeft aangetoond dat soortgelijke wijzigingen tussen 2012 en 2015 de score van Nederland hebben beïnvloed. Ook is het toetsingskader voor het domein lezen aangepast – met meer aandacht voor kritisch evalueren, waar Nederland relatief minder goed op presteert. Daarnaast liet onderzoek van Cito (2021) zien dat een correctie voor DIF in het peiljaar 2018 zou hebben geleid tot een hogere leesscore. Tot slot zagen we in 2018 een opvallend grote toename in het gebruik van één-uurs-boekjes en haalde Nederland het streefpercentage voor deelname van scholen niet. Het is onduidelijk wat per saldo het effect van al deze factoren op de positie van Nederland is geweest. We kunnen wel stellen dat de positie van Nederland hierdoor met meer dan de gebruikelijke onzekerheid is omgeven.

Een koppeling van de resultaten van peilingsonderzoeken aan onderwijsuitkomsten kan belangrijke inzichten geven in de ernst en aard van equivalentieproblemen voor Nederland. Op dit moment is het niet mogelijk om de data van PISA, TIMSS en PIRLS te koppelen aan de onderwijsuitkomsten van dezelfde leerlingen, zoals bijvoorbeeld de eindtoets, het leerlingvolgsysteem, en het eindexamen. Het koppelen van deze gegevens kan inzicht verschaffen in de representativiteit van de steekproef, die zeker in Nederland vanwege de hoge non-respons en het gebruik van één-uurs-boekjes een belangrijk aandachtspunt is. Een vergelijking van de uitkomsten van de peilingsonderzoeken met de resultaten van de eindtoets of eindexamens kan daarnaast inzicht geven in de mate waarin motivatie een rol speelt.

Hoewel equivalentieproblemen deels onontkoombaar zijn, valt er nog wel wat te winnen in het transparanter maken van de problemen en de gemaakte keuzes. Het zou goed zijn als de organisaties achter de peilingsonderzoeken meer inzicht geven in de gevolgen van belangrijke keuzes die onvermijdelijk enigszins arbitrair zijn. Hierbij valt onder andere te denken aan de selectie van ankervragen – waarmee ze corrigeren voor de moeilijkheidsgraad en de overstap naar digitaal en adaptief toetsen – en aannames bij de statistische modellering. Daarnaast kunnen de organisaties achter de internationale peilingsonderzoeken meer doen om de invloed van wijzigingen in het toetsingskader inzichtelijk te maken. Bijvoorbeeld door ook resultaten te publiceren die corrigeren voor een focusverschuiving. Op andere gebieden gebeurt dit al: de appendix van PISA 2018 bevat de cijfers gecorrigeerd voor veranderingen in de leerlingpopulatie en statistische methoden.

Literatuur

- Akyol, S.P., K. Krishna en 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 en 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 en 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 en R. Daugherty, 2011, *Policy effects of PISA*, Bristol: University of Bristol.
- Borghans, L. en T. Schils, 2012, *The leaning tower of Pisa: decomposing achievement test scores into cognitive and noncognitive components*, Working Paper, Maastricht: Maastricht University.
- Borgonovi, F. en 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 en 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 en C. Monseur, 2009, Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments, *IERI*, vol. 2, pag. 63-83.
- Hamhuis, E., C. Glas en 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 en 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.

Inspectie van het Onderwijs, 2020, *De Staat van het Onderwijs 2020*, Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs, 2021, *De Staat van het Onderwijs 2021*, Utrecht: Inspectie van het Onderwijs.

Jerrim, J., J. Micklewright, J.H. Heine, C. Salzer en 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 en 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. en 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 en L. Weijn, 2020, *Leerlingprestaties in de exacte vakken in groep 6 van het basisonderwijs: Resultaten TIMSS-2019*, Twente: Universiteit Twente.

Micklewright, J., S.V. Schnepf en 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 en M. van Onna, 2019, *Vaardigheidsontwikkelingen volgens PISA en exams*, 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. en 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*, Den Haag: Ministerie van Onderwijs, Cultuur en Wetenschap.

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

OESO, 2017, *PISA 2015 Technical report*, Parijs: OECD Publishing.

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

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

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

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

Onderwijsraad, 2022, *Taal en rekenen in het vizier*, Den Haag: Onderwijsraad.

Penk, C., C. Pöhlmann en 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. en 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 en 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.