VO

Vaardigheids-ontwikkelingen volgens PISA en examens

CITO

Vaardigheidsontwikkelingen volgens PISA en examens

Paul van der Molen, Sanneke Schouwstra, Remco Feskens en Marieke van Onna

Cito | Arnhem

Colofon

Aan dit rapport hebben ook meegewerkt:
Ger Limpens, Anneke de Graaf, Pieter Smeets, Ruud Stolwijk, Alex van de Kerkhof en Judith Veldhuizen.

Opmaak: Media Support
Foto omslag: Gijs Versteeg

Inhoud

Voorwoord

Naar aanleiding van het debat over examens in februari 2018 dienden Kamerlid Paul van Meenen c.s. een motie in over de vaardigheidsontwikkeling van leerlingen. In de motie stond de vraag centraal hoe het komt dat de prestaties van Nederlandse leerlingen bij het PISA-onderzoek een dalende tendens laten zien, terwijl tegelijkertijd de examencijfers gelijk blijven of zelfs stijgen. In verband daarmee werd gevraagd of de N-term en eventueel andere factoren het beeld op de vaardigheidsontwikkeling bevorderen of bemoeilijken.

Tijdens het debat over de motie werd duidelijk dat de Kamer er belang aan hechtte dat een onafhankelijke partij antwoord zou geven op de vragen in de motie. De Minister gaf aan dat hij het wenselijk achtte dat Cito en CvTE betrokken werden bij het proces van beantwoording. De minister besloot dat een onafhankelijke commissie antwoord zou geven op de vragen uit de motie. Cito en CvTE, maar ook andere partijen, kregen daarbij de gelegenheid om de commissie van de benodigde (technische) informatie te voorzien.

Om uitvoering te geven aan dit besluit deed OCW aan het Nationaal Regieorgaan Onderwijs-onderzoek (NRO) het verzoek om een onafhankelijke en deskundige commissie samen te stellen. Het NRO en OCW stemden de randvoorwaarden ten aanzien van het proces af, zoals de tijds-planning. NRO nam het secretariaat van de commissie op zich.

De commissie, onder voorzitterschap van Rick Steur, kwam in overleg met Cito overeen welke informatie Cito aan de commissie zou leveren. Een onderzoeksteam van Cito verzamelde vervolgens de benodigde gegevens, deed aanvullende analyses en schreef het rapport dat nu voor u ligt. Daarbij is – bijvoorbeeld over de passages over de normering – afstemming geweest met CvTE.

We willen de commissie Steur en CvTE bedanken voor de constructieve samenwerking bij de totstandkoming van dit rapport. Wij stellen het als Cito bijzonder op prijs dat we de gelegenheid hebben gekregen om vanuit onze expertise een bijdrage te leveren aan dit onderwijsdebat. Voor het onderwijs van vandaag en (over)morgen is het belangrijk om inzicht te hebben in de vaardigheidsontwikkeling van vo-leerlingen en de factoren die daarbij een rol spelen. Mocht op basis van dit advies verder technisch onderzoek nodig zijn, dan ondersteunen we daarin graag.

Cito
Arnhem
maart 2019

Samenvatting

Cito schreef dit technische rapport voor de commissie Steur. Deze onafhankelijke commissie werd geïnstalleerd naar aanleiding van een motie van Tweede Kamerlid Paul van Meenen c.s.. In de motie stond de vraag centraal waarom de examencijfers van Nederlandse leerlingen gelijk blijven of zelfs stijgen terwijl het PISA-onderzoek een dalende tendens laat zien. De motie vraagt om een nadere analyse en duiding van de vaardigheidsontwikkeling van leerlingen in Nederland. De commissie stuurt een brief aan de Minister waarin zij antwoord geeft op deze vraag.

Ontwikkelingen in prestaties en toetsing
De trends bij PISA zijn verschillend per domein. Terwijl leesvaardigheid vrijwel gelijk blijft op alle niveaus, dalen de prestaties op wiskundige geletterdheid geleidelijk tussen 2003 en 2015. Natuurwetenschappelijke geletterdheid blijft tot en met 2012 gelijk om in 2015 te dalen. Het gemiddelde van de OESO-landen laat ook een daling zien, maar in mindere mate. Ook de toetsing in PISA bleef niet ongewijzigd. Er vonden meerdere vernieuwingen plaats. Zo stapte PISA in 2015 over op een digitale afname met innovatieve opgaven. Ook werden andere analysemethoden geïntroduceerd.

Wat opvalt aan de examencijfers is dat de dalende trend in 2012 – met de aanscherping van de exameneisen – verandert in een stijgende trend, met name bij de kernvakken wiskunde en Engels. Ook bij de schoolexamens zien we dat de aanscherping van de exameneisen het beoogde effect heeft gehad: de daling werd omgezet in een stijging. De stijging van de schoolexamencijfers is wel kleiner dan van de cijfers op het centrale examen. Overigens heeft de stijging zich niet bij alle vakken en op alle niveaus doorgezet. Waar bijvoorbeeld bij Engels de stijging zich wel heeft doorgezet, zien we bij Nederlands voor havo/vwo een dip in het jaar van de ijking aan de referentieniveaus (2015). Bij wiskunde zien we weer een daling in 2016 en in 2017 bij havo, vmbo gl/tl en vmbo kb.

Normhandhaving
Het onderzoek belicht ook de normering van de centrale examens. In Nederland willen we dat de eisen die we aan leerlingen stellen, van jaar tot jaar gelijk zijn. De N-term corrigeert daarbij voor de moeilijkheid van een examen. Bij vakken waar normhandhavingsonderzoeken plaatsvinden, blijkt het goed mogelijk om de norm nauwkeurig te handhaven. Deze onderzoeken vinden vooral plaats bij vakken met veel kandidaten, zoals de meeste exacte vakken, de moderne vreemde talen en de digitale examens. Vakken waarbij geen normhandhavings-onderzoeken plaatsvinden (met name de vakken uit het cluster Mens en maatschappij en de kunstvakken) kunnen om inhoudelijke redenen slechts genormeerd worden in lijn met de normering van aanverwante vakken.

Vergelijking type onderzoek
Als we kijken naar het PISA-onderzoek en de Nederlandse centrale examens vallen vooral de verschillen op. Zo toetst PISA of leerlingen hun kennis en vaardigheden op het gebied van taal, wiskunde en natuurwetenschappen kunnen toepassen in alledaagse situaties, terwijl de centrale examens kennis en vaardigheden toetsen op een aantal duidelijk omschreven en onderwezen (technische) vakonderdelen. Bovendien wordt de PISA-toets afgenomen onder 15-jarigen, terwijl leerlingen bij het maken van de centrale examens meestal 16 tot 19 jaar oud zijn. Ook hebben de PISA-toetsen voor leerlingen geen consequenties, terwijl de centrale examens in Nederland hét meetmoment zijn voor doorstroom naar het vervolgonderwijs.

Leerlingenstromen
Dit onderzoek wijst ook uit dat de leerlingenpopulatie die start aan het eindexamen in de aanloop naar 2012 (aanscherping exameneisen) is veranderd. Zittenblijven en afstroom in de bovenbouw hebben vertraagd een effect op de examenprestaties. In de aanloop naar 2012 stegen het zittenblijven en de afstroom, wat samenhangt met een stijging van de eindexamencijfers aan het einde van het daarop volgende schooljaar. Deze veranderende leerlingenstromen kunnen bijvoorbeeld zijn veroorzaakt, doordat scholen zwaardere eisen zijn gaan stellen aan de bevordering naar een volgend leerjaar. Een stijging van het gemiddelde niveau van examenkandidaten is daarbij niet verrassend. Vanaf 2012 is een daling ingezet in zittenblijven en de afstroom, wat een verdere stijging van de eindexamencijfers in de daarop volgende jaren kan hebben afgezwakt. Opvallend is de geleidelijke toename in de periode 2015 – 2017 van het aantal leerlingen in het eindexamenjaar dat niet wordt aangemeld voor het eindexamen.

Belang en motivatie
Tot slot is gekeken naar het belang van de PISA-toets en de centrale examens. Voor PISA zijn er sterke aanwijzingen dat Nederlandse leerlingen minder gemotiveerd zijn dan leerlingen in andere landen. Nederlandse leerlingen antwoorden zeer snel en hun prestatie loopt gaandeweg de toets sneller terug dan in andere landen. Of de motivatie van Nederlandse leerlingen de afgelopen 10 jaar is afgenomen, is niet duidelijk.
Duidelijk is wel dat de motivatie van leerlingen om goed te presteren op de centrale examens sinds 2012 is toegenomen. De strengere diplomaeisen hebben het belang van de centrale examens verzwaard: voor zowel scholen als voor leerlingen. Als gevolg daarvan richten scholen zich meer op de centrale examens en bereiden ze leerlingen gerichter voor. Leerlingen zelf volgen vaker een georganiseerde examentraining.

Conclusie
Hoewel verwacht mag worden dat er een relatie bestaat tussen de prestaties van leerlingen op de PISA-toets en hun prestaties op de centrale examens, kan dit niet direct worden aangetoond. Dit rapport laat zien dat – op basis van alle beschikbare gegevens – de vergelijking van het PISA-onderzoek met de centrale examens, mank gaat. De voorbereiding op de toets, de inhoud van de toets en de leeftijd waarop leerlingen de toets maken, zijn verschillend. Door verzwaarde exameneisen is het belang van een goede score op de centrale examens de afgelopen jaren flink toegenomen. Leerlingen besteden meer aandacht aan de voorbereiding op de centrale examens en scholen besteden meer aandacht aan de stof van het centrale examen. Daarom is het niet verrassend dat trends bij de centrale examens anders zijn dan die bij PISA.

1 Inleiding

1 Inleiding

De motie van het lid Van Meenen c.s. van 13 februari 2018 luidde: “De Kamer, gehoord de beraadslaging, constaterende dat de prestaties van Nederlandse scholieren in het voortgezet onderwijs bij het internationale PISA-onderzoek een voortdurend dalende tendens laten zien op alle niveaus; constaterende dat de ontwikkeling van de gemiddelde cijfers voor het centraal examen gelijkblijvend tot stijgend is, in het bijzonder voor kernvakken sinds de introductie van de kernvakkenregeling, terwijl de ontwikkeling van de gemiddelde cijfers voor het school-examen gelijkblijvend is; overwegende dat voor de vaststelling van de cijfers van het centraal examen de N-termen een cruciale rol spelen; van mening dat het van groot belang is dat de kwaliteit van examens en de ontwikkeling van de vaardigheden van scholieren optimaal kunnen worden vastgesteld en gevolgd; verzoekt de regering, onafhankelijk onderzoek te laten uitvoeren naar de reële ontwikkeling van de vaardigheden van de scholieren in de afgelopen tien jaar, ten minste in de vakken wiskunde en Engels, en naar de mate waarin het gebruik van de N-termen en mogelijke andere factoren het beeld op deze ontwikkeling bevorderen of bemoeilijken, en de Kamer daarover voor 2019 te informeren en gaat over tot de orde van de dag.” (Van Meenen, Bruins, Van Haga, & Rog, Tweede Kamer, vergaderjaar 2017-2018, 31 289, nr. 359).

Deze motie vraagt om nader onderzoek. Dit onderzoek moet de reële vaardigheidsontwikkeling van scholieren in de afgelopen 10 jaar in beeld brengen, ten minste voor de vakken wiskunde en Engels. Belangrijke punten bij dit onderzoek zijn de factoren die het zicht op deze vaardigheids-ontwikkeling (kunnen) beïnvloeden en de resultaten (kunnen) vertekenen.

Dit rapport is een uitwerking van het onderzoek door Cito. Sommige begrippen in dit rapport kunnen technisch overkomen. Wie meer wil weten over de gebruikte begrippen, verwijzen we naar de begrippenlijst bij dit rapport en naar de toetsspecial over normering van examens[1]. Voor meer informatie over de opbouw van het voortgezet onderwijs (vo) in Nederland en de niveaus van examinering, verwijzen wij naar informatie op internet[2]. Dit eerste hoofdstuk start met een schets van de scope van het onderzoek. Daarna volgt de opzet van het onderzoek. We sluiten af met de onderzoeksvragen per hoofdstuk.

1.1 Scope van het onderzoek

Nederland neemt elke drie jaar deel aan het internationale peilingsonderzoek PISA. In het PISA-onderzoek worden vaardigheden van leerlingen gemeten op het gebied van wiskundige geletterdheid, natuurkundige geletterdheid en leesvaardigheid in de moedertaal. Dit onderzoek vindt plaats bij steekproeven van 15-jarige leerlingen. Doordat dit peilingsonderzoek om de drie jaar plaatsvindt, brengt dit vaardigheidsontwikkelingen in beeld. In Nederland zijn er geen nationale peilingsonderzoeken in het voortgezet onderwijs. Wel zijn er periodieke peilingen van het onderwijsniveau (PPON) in het primair onderwijs.

Als we in Nederland de vaardigheidsontwikkeling van vo-leerlingen in beeld willen brengen, kunnen we daarvoor – met enkele slagen om de arm – examenresultaten gebruiken.

1 http://www.toetsspecials.nl/html/normering/default.shtm

2 https://nl.wikipedia.org/wiki/Onderwijs_in_Nederland
https://www.govmbo.nl/voorgezet-onderwijs

In Nederland bestaan de eindexamens uit een schoolexamen en een centraal deel. In beide delen worden cijfers tussen 1 en 10 gebruikt voor de diploma-beslissing. Bijna alle leerlingen in Nederland doen eindexamen tussen 16 en 19 jaar, op één van de vijf examenniveaus. Dit betreft dan kernvakken als Nederlands, Engels en wiskunde, maar ook een tiental overige algemene vakken en in het vmbo ook beroepsgerichte examens.

Cijfers zijn niet direct vaardigheidsmetingen. Om de relatie tussen cijfers en vaardigheden te leggen gebruiken we normhandhavingsonderzoek bij de centrale eindexamens. Ieder jaar gebruiken we dit normhandhavingsonderzoek om de N-termen vast te stellen. N-termen bepalen de omzetting van score naar cijfer. Dezelfde gegevens van dit normhandhavings-onderzoek analyseren we voor dit rapport in een meerjarig perspectief. Zo brengen we vaardigheidsontwikkeling nogmaals in beeld.

In dit onderzoek diepen we de diverse gegevensbronnen nader uit en vergelijken we die met elkaar. Daarbij hebben we ons bij eindexamens expliciet beperkt tot de periode 2008 – 2017, bij PISA gebruiken we alle beschikbare momenten in de periode 2003 – 2015. Dit rapport focust op vakken die zowel in PISA als in de examens een rol spelen: Nederlands, wiskunde en natuur-wetenschappen. Daarbij belichten we voor natuurwetenschappen met name de vakken natuurkunde en biologie. Omdat de motie specifiek vraagt naar de vaardigheidsontwikkeling bij het vak Engels, voegen we ook Engels toe aan de lijst van te onderzoeken vakken, ondanks dat het geen plaats heeft in PISA.

1.2 Opzet van het onderzoek

Dit rapport begint in hoofdstuk 2 en 3 met een beschrijving van de trends in PISA en eindexamens.
In hoofdstuk 2 introduceren we de PISA-werkwijze en presenteren we de PISA-resultaten. Ook bespreken we factoren die bij die resultaten een rol speelden. In hoofdstuk 2 tonen we eveneens de ontwikkeling van examencijfers. Dit geldt voor zowel schoolexamencijfers als centraal examencijfers. Ook hiervan worden de factoren die van invloed zijn, besproken.
Hoofdstuk 3 start met een beschrijving van de normhandhavingswijze van de Nederlandse eindexamens in vo. De resultaten van vakken met ankeropgaven worden gepresenteerd in een aantal grafieken. We geven per vak de geschatte vaardigheidsontwikkeling samen met de daadwerkelijke cijferontwikkeling weer.

Daarna bespreken we in hoofdstuk 4 t/m 6 enkele mogelijke verklaringen voor de geconstateerde verschillen in trends in PISA en eindexamens.
In hoofdstuk 4 beschrijven we de inhoudelijke verschillen en overeenkomsten tussen de eindtermen van het examenprogramma en de domeinbeschrijvingen van PISA. Ook laten we zien hoe deze inhoudsbeschrijvingen uitgewerkt worden in opgaven.
In hoofdstuk 5 gaan we in op de samenstelling van de onderzochte populaties. Bij PISA wordt standaard gewerkt met een representatieve steekproef. Bij eindexamens is dit niet het geval. We gaan daarom in op veranderingen in leerlingenstromen bij de centrale examens.
In hoofdstuk 6 besteden we aandacht aan de invloed van motivatie op gemeten prestaties. We onderzoeken hoe groot de motivatie is van Nederlandse leerlingen tijdens de PISA-afnames vergeleken met andere landen. Bij de centrale examens onderzoeken we of de aanscherping van de exameneisen in 2012 en 2013 een verandering teweeg heeft gebracht in de prestaties van de leerlingen.

We sluiten af met een slotbeschouwing in hoofdstuk 7. Er volgt een samenvatting van de onderzoeksresultaten en we geven enkele beperkingen van het onderzoek aan. Tot slot beschrijven we onze kijk op de geconstateerde verschillen tussen trends binnen PISA en eindexamens in vo.

1.3 De onderzoeksvragen per hoofdstuk

In dit rapport geven we per hoofdstuk antwoord op diverse onderzoeksvragen. Een overzicht:

- Hoofdstuk 2: Hoe zag de vaardigheidsontwikkeling op de domeinen taal, wiskundige geletterdheid en natuurwetenschappelijke geletterdheid eruit in PISA-onderzoeken in de periode 2006 – 2015? Welke factoren waren hier van invloed op de waargenomen ontwikkelingen?
  Hoe zag de cijferontwikkeling op het centraal examen en het schoolexamen eruit voor de relevante vakken (zie 1.3: De scope van dit onderzoek) in de afgelopen 10 jaar? Hoe waren factoren als gewijzigde examenprogramma's en de aangescherpte eisen in de zakslaagregeling van invloed op de waargenomen ontwikkelingen?
- Hoofdstuk 3: Voor welke vakken is normhandhavingsonderzoek uitgevoerd en voor welke vakken niet, en waarom niet? Hoe komt de technische N-term tot stand? Hoe komt de definitieve N-term tot stand en hoe verhoudt deze zich tot de technische N-term? Hoe zag de vaardigheidsontwikkeling eruit bij de vakken die in dit onderzoek centraal staan?
- Hoofdstuk 4: Welke verschillen en overeenkomsten bestaan er tussen de domeinbeschrijvingen van PISA en de eindtermen van het examenprogramma? Hoe worden deze inhoudsbeschrijvingen geoperationaliseerd in de opgaven?
- Hoofdstuk 5: Welke invloed heeft een wijzigende leerlingenstroom gehad op de examenresultaten?
- Hoofdstuk 6: Wat is de invloed van de examens op de examenresultaten? Welke invloed heeft motivatie op het prestatieniveau zoals gemeten in de centrale examens en in PISA?

2 De gepubliceerde cijfers

2 De gepubliceerde cijfers

Hoe zien de prestaties van Nederlandse leerlingen de afgelopen tien jaar eruit, als we kijken naar PISA en naar de centrale examens? Welke trends en ontwikkelingen zien we? Zijn er wijzigingen in het PISA-onderzoek en in de examenregeling die in verband gebracht kunnen worden met deze trends? In dit hoofdstuk beschrijven we de resultaten van PISA en die van de schriftelijke centrale examens.

Aan de hand van PISA-rapporten onderzoeken we de vaardigheidsontwikkeling van Nederlandse scholieren. We doen dat voor de gemeten vaardigheden 'leesvaardigheid (in de moedertaal)', natuurwetenschappen en wiskunde, in de periode 2003 – 2015 (vijf afnames[3]). We gebruiken gegevens van de OESO om trends in de resultaten te beschrijven. Daarbij vergelijken we de vaardigheid van Nederlandse scholieren met het OESO-gemiddelde en bekijken we de vaardigheidstrend per schooltype en leerweg (vmbo-bb, vmbo-kb, vmbo-gt, havo en vwo). Voor een zuiver inzicht geven we aan hoe het PISA-onderzoek is uitgevoerd en welke vernieuwingen en wijzigingen in de loop der jaren plaatsvonden.

De ontwikkeling van examencijfers beschrijven we aan de hand van de gegevens uit de Examenmonitor (DUO). We doen dat voor de afgelopen 10 jaar (2008-2017), zowel voor cijfers van de centrale examens, als de schoolexamens. We kijken of trends te relateren zijn aan wijzigingen in de examenregeling (zoals de regel dat het CE-cijfer gemiddeld minstens een 5,5 moet zijn, of de kernvakkenregeling).

2.1 Trends in PISA

Vaardigheden

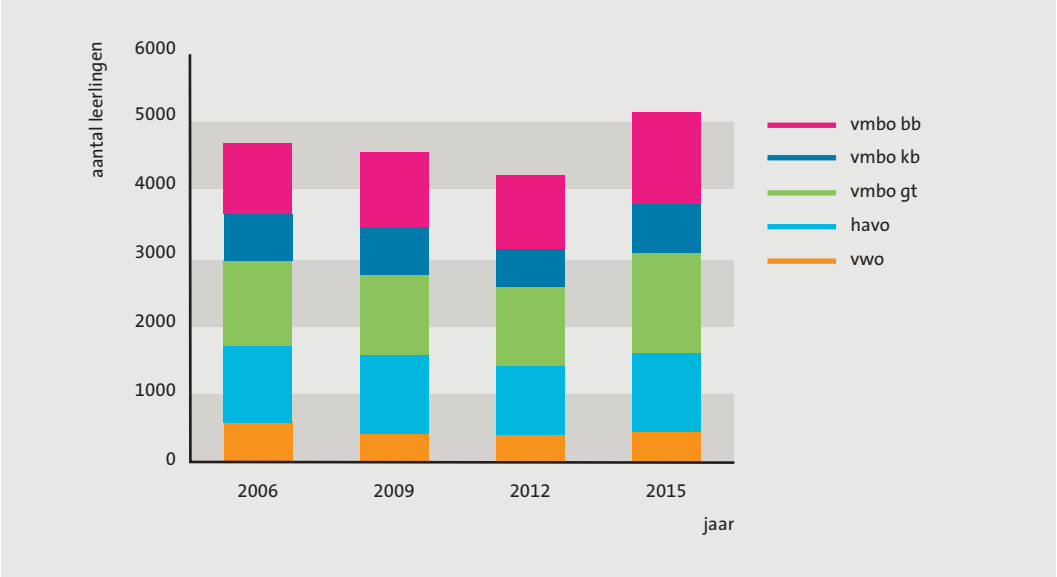
PISA is een internationaal vergelijkend onderzoek naar de kennis en vaardigheden van 15-jarige leerlingen wereldwijd. PISA evalueert in hoeverre 15-jarigen hun kennis en vaardigheden kunnen toepassen in alledaagse situaties. PISA toetst niet wat leerlingen precies hebben geleerd op school, maar meet het vermogen van leerlingen om realistische taken te vervullen die kernvaardigheden vereisen. Dit wordt "geletterdheid" genoemd. Sinds 2000 wordt PISA iedere drie jaar uitgevoerd onder leiding van de OESO met een wisselend internationaal consortium van onderzoeksinstituten. In elke afname ligt de nadruk op één van de drie hoofddomeinen (leesvaardigheid, wiskundige geletterdheid of natuurwetenschappelijke geletterdheid). In 2000 en 2009 was leesvaardigheid het centrale domein, in 2003 en 2012 wiskundige geletterdheid en in 2006 en 2015 stond natuurwetenschappelijke geletterdheid centraal. Leerlingen maken alleen van het centrale hoofddomein veel vragen en er wordt meer diepgaande informatie verzamelt over het centrale hoofddomein. Bijvoorbeeld in 2015 werd ook de houding van leerlingen ten opzichte van natuurwetenschappelijke geletterdheid en hun opvattingen over natuurwetenschappen onderzocht.

3 PISA 2018 is niet opgenomen, omdat over de afname van 2018 nog geen rapporten zijn verschenen. Het internationale rapport van OESO over PISA 2018 zal eind 2019 verschijnen.

Aantal deelnemers
Steeds meer landen en leerlingen doen mee aan het PISA-onderzoek. In 2000 deden ruim 250.000 leerlingen uit 32 landen mee (waarvan 27 OESO-landen). In 2015 was het aantal leerlingen gestegen naar circa 540.000 uit 71 verschillende landen en jurisdicties. Alle 35 landen die lid zijn van de OESO en 36 niet-lidstaten (de zogenaamde partnerlanden) participeerden in 2015 in PISA.

In Nederland doen elke keer zo'n 5.000 15-jarige leerlingen mee aan een PISA-afname (zie figuur 2.1). Deze leerlingen zitten vrijwel allemaal in leerjaar 3 (42% in 2015) of leerjaar 4 (55% in 2015). Dat betekent dat deelnemende vmbo-leerlingen hetzelfde jaar of een jaar later nog eindexamen doen. Bijvoorbeeld in de PISA-afname van 2015 (zie Cito, 2016, p. 23) zat bijna de helft van de 2649 vmbo-leerlingen (vmbo-bb, vmbo-kb en vmbo-gt) in het examenjaar (48,2%). Bijna alle havo-leerlingen zitten één tot twee jaar later in het examenjaar en vwo-leerlingen twee tot drie jaar later.

Deelname aan PISA is in Nederland voor scholen en leerlingen vrijwillig. In Nederland zijn voor leerlingen en scholen geen consequenties verbonden aan hoe goed de leerlingen het doen op de toetsen. Het internationale consortium trekt voor elk land een representatieve steekproef van 15-jarige leerlingen en controleert of de steekproef voldoet aan alle kwaliteitseisen, zoals voldoende respons. Feskens en Koops hebben in 2016 onderzoek gedaan naar een eventuele vertekening van de resultaten ten gevolge van nonrespons in Nederland en zij konden geen afwijkingen van representativiteit in Nederland vaststellen.



Figuur 2.1 Aantal leerlingen in Nederland dat heeft deelgenomen aan het PISA-onderzoek

Opgaven
Het internationaal consortium beschrijft de vaardigheden die getoetst worden, en de toetsen zelf. In de jaren dat een domein centraal staat, wordt de omschrijving van die vaardigheid en opgaven vernieuwd. Daarbij wordt ervoor gezorgd dat de vernieuwde toets altijd een aantal opgaven gemeenschappelijk heeft met de oorspronkelijke toets. Dit zorgt ervoor dat resultaten met elkaar vergeleken kunnen worden. De opgaven blijven grotendeels geheim, zodat toetsen meerdere malen afgenomen kunnen worden en resultaten via gemeenschappelijke opgaven in toetsen uit verschillende jaren (de ankeropgaven) met elkaar vergeleken kunnen worden.

Plausible values
Bij een PISA-afname maken leerlingen veel opgaven van het domein dat in die specifieke afname centraal staat. Daarnaast maken ze een kleinere selectie van opgaven uit de toetsen voor de andere twee hoofddomeinen. Omdat het aantal opgaven per leerling te gering is voor een nauwkeurige vaardigheidsschatting, worden in PISA zogenaamde "plausible values" (aannemelijke waarden) geschat.
Op grond van de antwoorden op de subset van items en beschikbare achtergrondinformatie wordt per leerling meerdere keren geschat wat een aannemelijke waarde is (Mislevy, 1991). Een dergelijke aannemelijke waarde kan zelfs worden geschat indien een leerling helemaal geen items van een bepaald domein heeft gemaakt (op grond van de antwoorden bij de andere domeinen). Aannemelijke waarden zijn geen traditionele individuele scores en kunnen daardoor ook niet gebruikt worden voor individuele rapportage, maar ze leveren wel betere[4] schattingen op van de vaardigheid en variantie in de (sub-)populatie (zie bijlage 2 over plausible values).

Voor de leerlingen worden dus geen traditionele individuele scores berekend, maar alleen aannemelijke waarden geschat. Daarnaast worden de gegevens van individuele leerlingen geheel geanonimiseerd. Mede door deze twee eigenschappen kunnen de PISA-resultaten niet op leerlingniveau gekoppeld worden aan andere resultaten, zoals de eindexamencijfers.

2.1.1 Wijzigingen in het PISA-onderzoek

Afname en opgaven
Tot 2012 was de afname van PISA uitsluitend op papier. In 2012 werd de mogelijkheid geboden om een deel van de wiskundeopgaven ook digitaal af te nemen. Nederland bleef echter – zoals de meeste landen – bij papier. In 2015 werd een digitale afname de PISA-standaard. Ook Nederland stapte toen over. Tegelijk met de introductie van de volledig digitale afname werden nieuwe typen opgaven, met interactieve animaties, gebruikt voor natuurwetenschappelijke geletterdheid. Bij deze opgaven moesten leerlingen variabelen manipuleren in gesimuleerd natuurwetenschappelijk onderzoek. Een andere wijziging was dat er zeer veel verschillende toetsversies waren (396, OECD, 2016, p. 22) en dat de mogelijkheid om terug te bladeren tijdens de afname beperkt was.

Analyse
In 2015 gebruikte PISA ook andere analysemodellen. PISA gebruikt IRT-modellen, zoals in Nederland bijvoorbeeld ook bij de Eindtoets Basisonderwijs en de Centrale Eindtoets gebeurde. In het PISA-onderzoek werd tot en met 2012 een minder complex IRT-model gebruikt. In 2015 werden daarnaast meer complexe IRT-modellen[5] gebruikt vanwege zorgen over de geschiktheid van het eerdere model (OECD, 2017, p. 142). Ook de procedure om scores van verschillende jaren te vergelijken (equivalering), werd gewijzigd. Men kan twee procedures onderscheiden: gelijkstellen of koppelen. Tot en met 2012 gebruikte PISA gelijkstellen (waarbij statische vraageigenschappen per afname bepaald worden en op dezelfde schaal geplaatst worden met behulp van een lineaire transformatie). Sinds 2015 wordt koppelen gebruikt (waarbij de schaal wordt vastgelegd met behulp van vragen die beide afnames gemeenschappelijk hebben en alle beschikbare data uit beide afnames). Het is niet geheel uit te sluiten dat de andere methodes net iets andere resultaten zouden hebben opgeleverd.

4 Minder vertekening en dus naar verwachting dichter bij de werkelijke waarde

5 Tot 2015 werden itemparameters geschat met een zogenaamd Rasch-model (Rasch, 1960) en een partial credit model (Masters, 1982), maar in 2015 werden daarnaast meer complexere IRT-modellen gebruikt zoals het 2-parameter-logistisch model (Birnbaum, 1968) en een generalised partial credit model (Muraki, 1992).

Domeinen
Naast een toename van deelnemende landen en leerlingen, de overstap naar een digitale afname, nieuwe opgavevormen, andere afnamecondities en andere analyse- en equivaleringstechnieken, werden de toetsen zelf allemaal inhoudelijk vernieuwd. In 2015 is de definitie van natuurwetenschappelijke geletterdheid grondig herzien en zijn ook de subdomeinen van natuurwetenschappelijke geletterdheid gewijzigd.

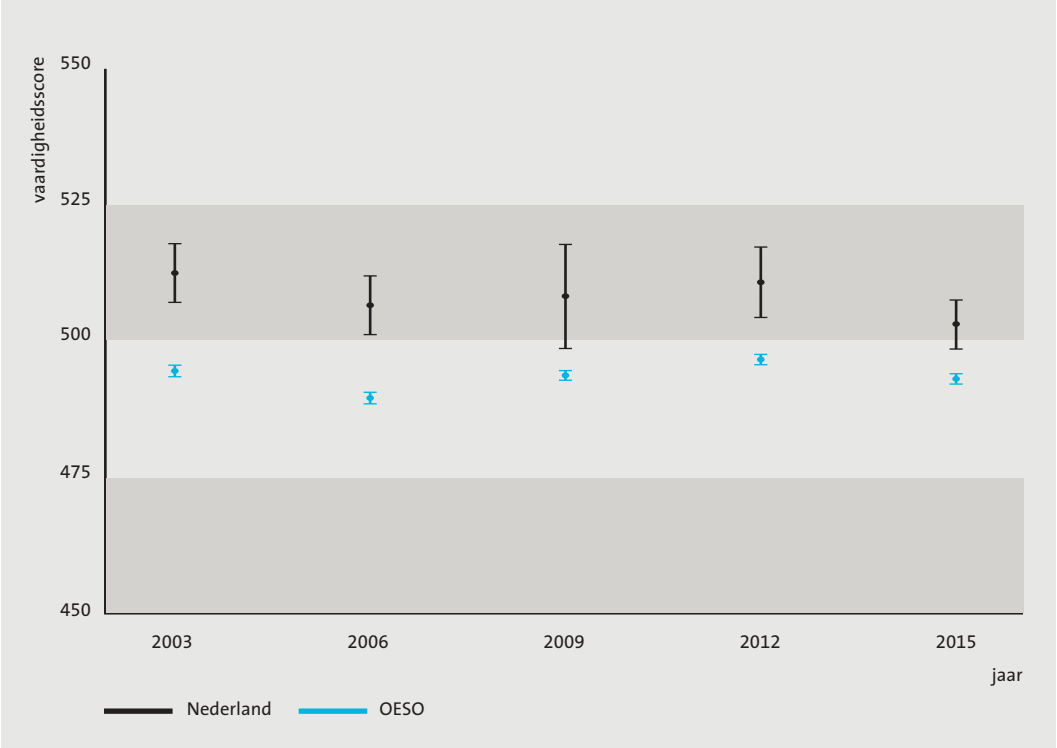
2.1.2 Trendresultaten: gemiddelden in Nederland

Nederland doet vanaf de eerste afname in 2000 mee aan het PISA-onderzoek. In 2000 was de respons alleen te laag om opgenomen te worden in de rapportage. De respons voldeed niet aan de minimumeis van de OESO. Daarom zijn de resultaten uit 2000 niet meegenomen in dit onderzoek. Bij elke latere afname is de PISA-toets gemaakt door ongeveer 5000 leerlingen in Nederland. Deze leerlingen zijn zodanig geselecteerd dat ze representatief zijn voor alle 15-jarigen.

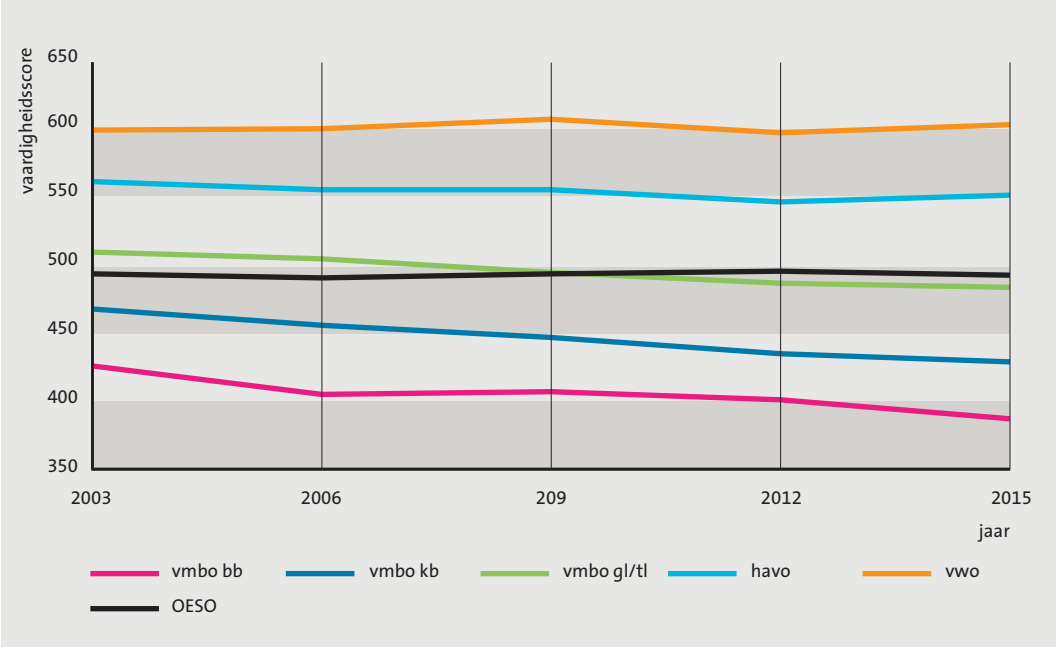
Deze paragraaf beschrijft de trendresultaten op de PISA-hoofddomeinen. De Nederlandse schattingen worden vergeleken met de gemiddelde vaardigheidsschatting van de OESO-landen (de landen die lid zijn van de Organisatie voor Economische Samenwerking en Ontwikkeling en meedoen aan het PISA-onderzoek). Alle resultaten van het PISA-onderzoek worden gepresenteerd op schalen die zijn gestandaardiseerd op een internationaal gemiddelde van 500, met een standaardafwijking van 100. Deze spreidingsmaat impliceert dat ongeveer twee derde deel van de leerlingen op een score tussen 400 en 600 uitkomt (500 ± 100). Het gemiddelde van 500 geldt alleen voor de OESO-landen en is vastgezet in het jaar dat het betreffende onderwerp voor het eerst hoofdonderdeel was. Dat wil zeggen in 2000 voor leesvaardigheid, in 2003 voor wiskundige geletterdheid en in 2006 voor natuurwetenschappelijke geletterdheid.
De resultaten van de partnerlanden die meedoen aan PISA worden afgezet tegen het gemiddelde van de OESO-landen.

Leesvaardigheid
Leesvaardigheid wordt in PISA gedefinieerd als ‘het begrijpen, gebruiken, reflecteren en betrokken zijn bij geschreven teksten om je doelen te bereiken, je kennis en potentieel te verruimen, en deel te nemen aan de maatschappij’. Figuur 2.2 toont de gemiddelde vaardigheidsscores voor leesvaardigheid in Nederland en de OESO-landen sinds 2003.
De verticale lijnen geven het 95% – betrouwbaarheidsinterval rondom het gemiddelde weer. Wat blijkt is dat de leesvaardigheid van Nederlandse 15-jarigen in de periode 2003-2015 niet wezenlijk is veranderd. Geen van de verschillen tussen opeenvolgende afnames is namelijk significant. In figuur 2.3 zijn de prestaties per opleidingstype weergegeven[6].

6 In de figuren met de prestaties per opleidingstype zijn geen betrouwbaarheidsintervallen weergegeven, omdat in de publicaties over de PISA-resultaten de standaardfouten per opleidingstype niet gegeven worden.



Figuur 2.2 De gemiddelde leesvaardigheid in Nederland en OESO-landen in het PISA-onderzoek



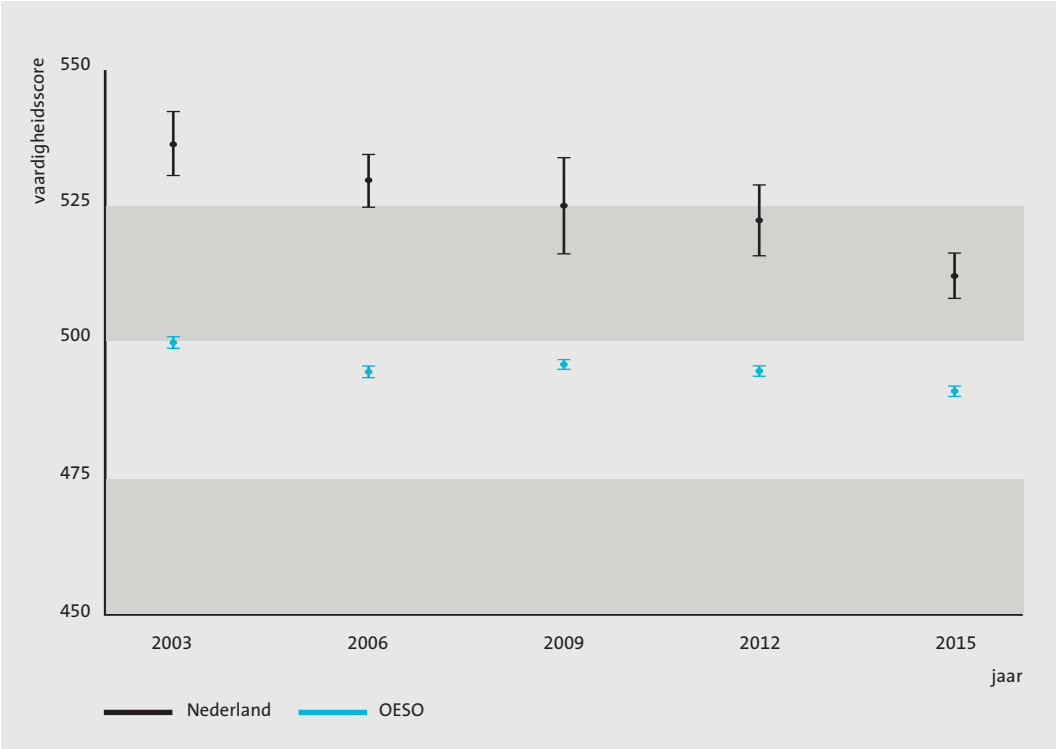
Figuur 2.3 Gemiddelde leesvaardigheid in Nederland per opleidingstype in het PISA-onderzoek

Wiskundige geletterdheid

Sinds 2012 geldt een nieuwe formele definitie van wiskundige geletterdheid. Wiskundige geletterdheid is 'het vermogen van een individu om wiskunde in een diversiteit van contexten te formuleren, gebruiken en interpreteren'. Naast de inhoudelijke deeltoetsen (voor Vorm en ruimte, Veranderingen en Relaties, Onzekerheid, Hoeveelheid), werden in 2012 drie nieuwe schalen ontwikkeld voor het meten van wiskundige competenties (formuleren, toepassen en interpreteren). De nieuwe wiskundetoets hield een aantal gemeenschappelijke opgaven met de

wiskundetoets uit 2003. Daardoor bleef het mogelijk om de scores van verschillende jaren op dezelfde schaal weer te geven.

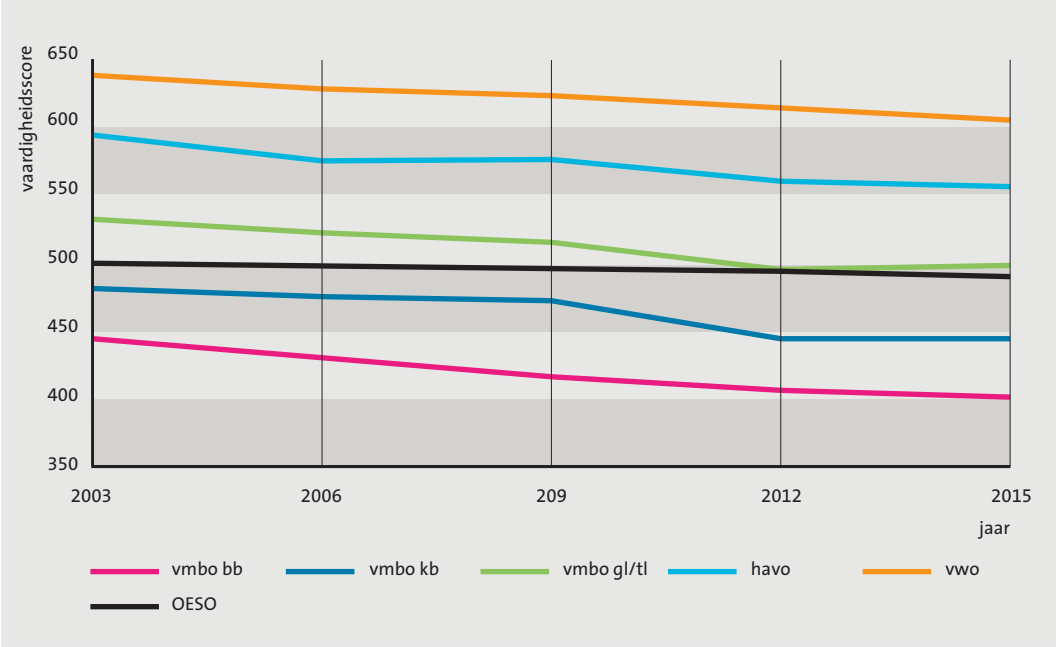
Figuur 2.4 toont de gemiddelde vaardigheidsscores voor wiskundige geletterdheid in Nederland en de OESO-landen. Zichtbaar is dat de gemiddelde score tot en met 2012 telkens licht daalt. In 2015 is de daling zelfs nog wat groter geworden. De daling tussen opeenvolgende metingen is nooit significant, maar in 2015 is de vaardigheidsscore wel significant lager dan in 2003 en 2006. Ook voor de OESO-landen als geheel gold een gemiddelde daling.



Figuur 2.4 Gemiddelde wiskundige geletterdheid in Nederland en OESO-landen in het PISA-onderzoek

Kijken we per opleidingstype (figuur 2.5), dan vindt in 2012 – met de vernieuwing van de toets wiskundige geletterdheid – bij alle opleidingstypes een lichte daling plaats. Tussen 2012 en 2015 is bij havo/vwo en vmbo-bb nog een hele lichte daling[7] te zien. Bij vmbo-gt en vmbo-kb bleven de prestaties gelijk.

7 In de publicaties over het PISA-onderzoek wordt de standaardfout en significantie per opleidingstype niet gegeven.



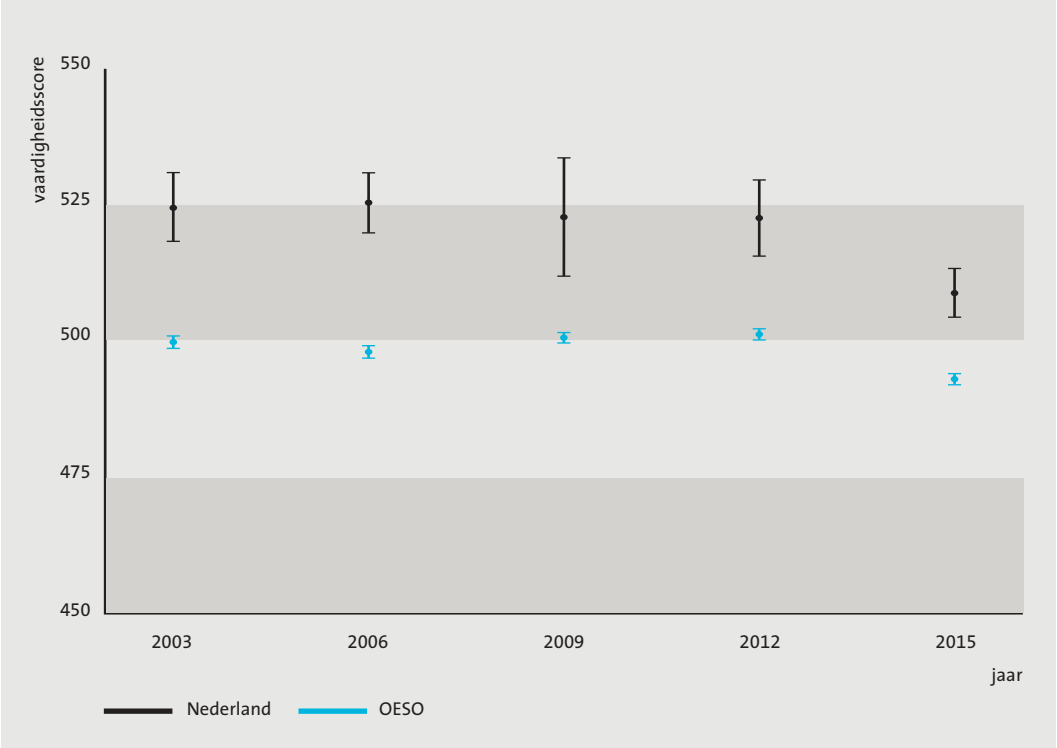
Figuur 2.5 Gemiddelde wiskundige geletterdheid in Nederland per opleidingstype in het PISA-onderzoek

Natuurwetenschappelijke geletterdheid

Natuurwetenschappelijke geletterdheid werd in 2015 geherdefinieerd als 'het vermogen om zich bezig te houden met natuurwetenschappelijke kwesties en met natuurwetenschappelijke ideeën als weldenkend burger. Een natuurwetenschappelijk geletterde persoon is bereid om een beredeneerd betoog over natuurwetenschap en technologie te houden.' In 2006 werden twee subdomeinen onderscheiden: 'Kennis van natuurwetenschappen' en 'Kennis over natuurwetenschappen'. In 2015 worden de twee subdomeinen Vakkennis en Kennisvorming gemeten. Daarbij wordt bij kennisvorming onderscheid gemaakt tussen procedurele en epistemische kennis. Deze inhoudelijke wijzigingen hebben tot gevolg dat er geen een-op-een relatie meer bestaat tussen de subdomeinen van natuurwetenschappelijke geletterdheid in 2015 en die van eerdere afnames. In hoofdstuk 4, de inhoudsanalyse, zal uitgebreider worden stilgestaan bij de definities en de opgaven.

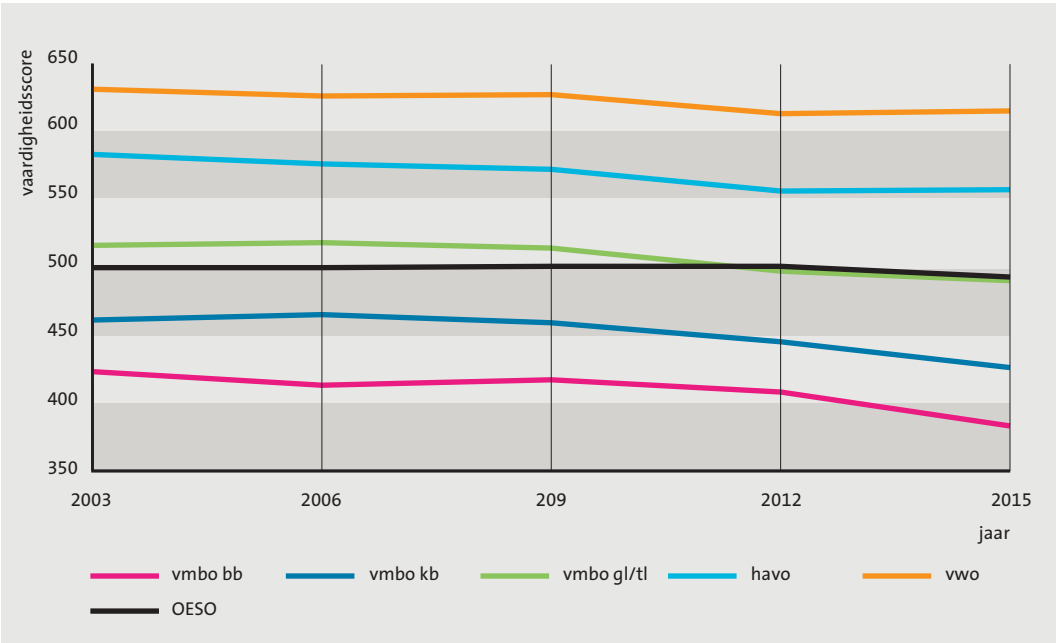
Figuur 2.6 toont de natuurwetenschappelijke geletterdheid van Nederlandse 15-jarigen in de periode 2003 – 2015. Die blijkt over de jaren heen niet wezenlijk veranderd. De verschillen tussen 2006, 2009 en 2012 zijn zeer klein en statistisch niet significant. Dit laatste is wel het geval voor de recente vergelijking van 2012 met 2015.
De ontwikkeling van het vaardigheidsniveau in Nederland lijkt op die in de OESO-landen. Ook in OESO-verband zijn de prestaties in de periode 2006 – 2012 relatief stabiel, om van 2012 naar 2015 te dalen. In dit jaar is natuurwetenschappelijke geletterdheid geherdefinieerd en zijn meerdere vernieuwingen doorgevoerd. In 2015 is overgestapt naar een digitale afname, nieuwe interactieve opgavevormen voor wetenschappelijke geletterdheid, andere afnamecondities en andere analyse- en equivaleringstechnieken.

Wel is de daling in Nederland sterker dan gemiddeld in de OESO-landen. Wat blijkt is dat de gemiddelde vaardigheid van Nederlandse 15-jarigen sinds 2012 met dertien punten is gedaald (tegenover acht punten in OESO-verband). Maken we een vergelijking met 2006, het moment dat natuurwetenschappelijke geletterdheid voor het eerst het hoofddomein was, dan zijn de prestaties van Nederlandse leerlingen met zestien scorepunten gedaald (tegenover zeven punten in OESO-verband).



Figuur 2.6 Gemiddelde natuurwetenschappelijke geletterdheid in Nederland en OESO-landen in het PISA-onderzoek

Bekijken we de trends per opleidingstype (schooltype/leerweg), dan toont figuur 2.7 dat de daling in 2015 alleen optreedt bij vmbo-leerlingen.



Figuur 2.7 Gemiddelde natuurwetenschappelijke geletterdheid in Nederland per opleidingstype in het PISA-onderzoek

2.2 Trends in de centrale examens

2.2.1 Wijzigingen in de examens en examenregeling

Stand van zaken 2007
Op het havo/vwo bestonden vier profielen waarbij veel vakken een heelvak/deelvakstructuur hadden. Dit hield in dat leerlingen die het deelvak deden, slechts een deel van het hele vak volgden. Ze deden ook een eigen examen. Leerlingen kozen bijvoorbeeld een deelvak, als dat vak deel uitmaakte van het gekozen profiel, maar volgens hen niet erg aantrekkelijk was.
Het behoeft weinig betoog dat leerlingen die een heelvak gevolgd hadden, gemiddeld genomen vaardiger waren dan leerlingen die een deelvak volgden.
De vakken met een heelvak/deelvakstructuur waren:

- Vwo: Duits, Frans, wiskunde A, wiskunde B, natuurkunde, scheikunde, biologie en economie.
- Havo: Duits, Frans, wiskunde A, **wiskunde B**, **natuurkunde**, **economie**, Fries, Spaans, Russisch, Turks, Arabisch.

Voor de vetgemaakte vakken kende het deelvak een centraal examen. De overige vakken sloten het deelvak af met alleen een schoolexamen.

In 2007 startten 4-havo en 4-vwo met een nieuwe indeling, de zogenaamde PEP[8]-operatie. Daarbij kwamen de heel- en deelvakken te vervallen. Ook werd examenstof opnieuw verdeeld over het centraal en schoolexamen. Het centraal examen besloeg niet langer alle stof. Bij veel vakken (wiskunde en natuurkunde uitgezonderd) werd 60% van het examenprogramma aan het centraal examen toegewezen, de overige 40% werd afgesloten met een schoolexamen.
Vmbo-bb kende vanaf 2005 digitale examens. Het aandeel scholen dat overging op deze afnamevorm groeide snel. In 5 jaar tijd was het aandeel meer dan 90%. In vmbo-kb was enkele jaren later eenzelfde patroon zichtbaar.

Ontwikkelingen in 2009/2010
In 2009 werden de eerste havo-examens volgens de nieuwe indeling afgenomen. In 2010 volgde het vwo. Het aantal leerlingen dat vakken met een oude heelvak/deelvakstructuur volgde, bleef redelijk constant. Toch was lastig te bepalen hoe de totale populatie in de nieuwe situatie presteerde. Niet alleen was sprake van het samenvoegen van leerlingen met meer of minder talent voor het vak, maar het onderwijs was ook behoorlijk veranderd. Door harmonisatie van onderwijstijd voor de nieuwe vakken, waren voor een vak als wiskunde minder studielasteenheden beschikbaar. Ook deden veel meer leerlingen eindexamen in veel vakken die eerst een heelvak/deelvakstructuur hadden. Over de vaardigheid van de oude deelvakkers was niet veel bekend; zij hadden hun vak immers afgesloten met een schoolexamen. Tot slot ging het deelvak vwo wiskunde A1 over in wiskunde C. Dit was het enige deelvak dat als volwaardig vak bleef bestaan in de nieuwe situatie.

Voor het vmbo waren er niet veel wijzigingen in deze periode. Het was de tijd van de start van het digitale examen in vmbo-kb.

Ontwikkelingen in 2011-2013
In 2012 werd de zak/slaagregeling aangescherpt met de CE-eis. Voor het eerst moesten kandidaten over alle vakken in het centraal examen gemiddeld een voldoende (5,50) halen. Afgaande op de cijfers zoals die in de jaren 2010-2011 waren behaald, zou dit een grote toename van het aantal gezakte kandidaten tot gevolg hebben. De aankondiging van de nieuwe

8 Platform Examen Programma's

uitslagregel kan op scholen geleid hebben tot kritischere bevorderingsnormen en bij leerlingen tot extra examenvoorbereiding. Immers, met hoge SE-cijfers was je nog niet zeker van een diploma.

In 2013 werd de zak/slaagregeling verder aangescherpt met de kernvakkenregeling. Havo/vwo-kandidaten mochten vanaf dat moment voor de kernvakken Nederlands, Engels en wiskunde niet meer dan één vijf halen. Op het vmbo moest vanaf 2014 het vak Nederlands afgesloten worden met minimaal een vijf.

2.2.2 Ontwikkelingen 2014 – nu

In deze periode waren er geen grote stelselwijzigingen. Wel werd in 2014 de rekentoets onderdeel van het eindexamen. Leerlingen moesten de rekentoets minimaal één keer afgelegd hebben voor diplomering. Het cijfer op de rekentoets telde op het vwo twee jaar mee in de zak/slaagbeslissing: in 2016 bepaalde de rekentoets mede of het diploma was behaald, in 2017 telde de rekentoets mee in de kernvakkenregeling (DUO, 2017, Examenblad, 2017).

Wijzigingen op vakniveau
Op vakniveau vonden in de periode 2007 – 2018 enkele grote wijzigingen plaats. Zo werd op havo/vwo vanaf september 2013 volgens een nieuw examenprogramma lesgegeven in de natuurwetenschappelijke vakken. Vanaf september 2015 werd ook voor wiskunde een nieuw examenprogramma van kracht. De eerste landelijke examens voor deze vakken waren:

- 2015 natuurkunde, scheikunde en biologie havo
- 2016 natuurkunde, scheikunde en biologie vwo
- 2017 wiskunde havo
- 2018 wiskunde vwo

Bij het vak Nederlands werden de examens met ingang van 2015 geijkt aan de referentieniveaus. Dit betekende onder meer dat op het vwo – om te voldoen aan de criteria van referentieniveau 4F – de norm 0,4 cijferpunt strenger werd. Bij vmbo-bb werd de norm 0,1 cijferpunt strenger en lag daarmee 1,0 cijferpunt onder referentieniveau 2F. De havo- en vwo-examens kenden in 2015 nog een andere, meer inhoudelijke, wijziging. De geleide samenvatting verviel. Met ingang van 2016 werd een correctie op taalfouten ingevoerd. Leerlingen die bij het beantwoorden van vragen meerdere taalfouten maakten, kregen aftrekpunten.

Tabel 2.1 Overzicht wijzigingen in de examens en examenregeling

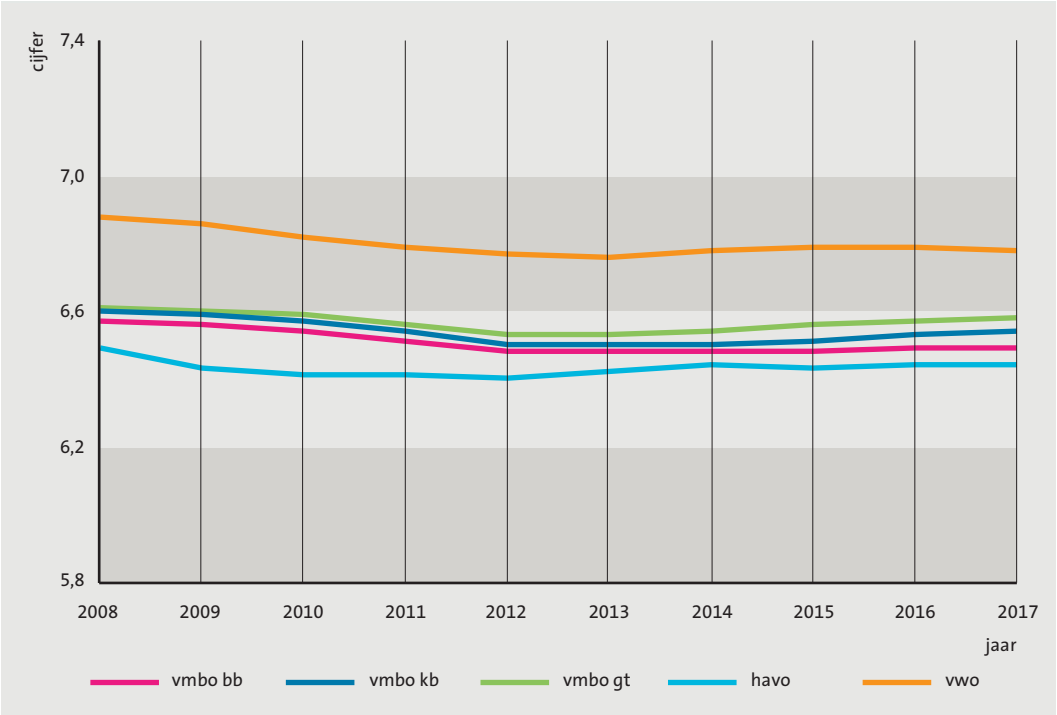
	vmbo	havo/vwo
2007	Vmbo-bb: toename digitale afnames	Examenprogramma voor 60% toegewezen aan CE (m.u.v. wiskunde en natuurwetenschappen)
2008		
2009		Havo: afschaffen heelvak/deelvak in CE
2010	Vmbo-kb: begin digitale afnames	Vwo: afschaffen heelvak/deelvak in CE
2011		
2012	Aanscherpen exameneis (> 5,5 gem)	Aanscherpen exameneis (> 5,5 gem)
2013		Kernvakkenregeling
2014	Rekentoets (resultaat telt niet mee in slaag/zak-regeling). Exameneis Nederlands (>5).	Rekentoets (resultaat telt niet mee in slaag/zak-regeling).
2015	CE Nederlands geijkt aan referentieniveaus. Rekentoets (resultaat telt niet mee in slaag/zak-regeling).	Havo: nieuw examenprogramma natuurwetenschappen. CE Nederlands geijkt aan referentieniveaus; afschaffing geleide samenvatting in CE Nederlands. Rekentoets (resultaat telt niet mee in slaag/zak-regeling).
2016	Rekentoets (resultaat telt niet mee in slaag/zak-regeling).	Vwo: nieuw examenprogramma natuurwetenschappen. Vwo: rekentoets telt mee in examenresultaat.
2017		Havo: nieuw examenprogramma wiskunde. Vwo: rekentoets telt mee in kernvakkenregel.
2018		Vwo: nieuw examenprogramma wiskunde

2.2.3 Trends in het gemiddelde cijfer op het schoolexamen en het centrale examen

Jaarlijks publiceert DUO de Examenmonitor. Die geeft de trends bij de kernvakken weer. We hebben de ontwikkelingen in de periode 2008-2017 bekeken. Daarvoor hebben we de gegevens uit de Examenmonitor 2013 samengevoegd met de gegevens uit de Examenmonitor VO 2017.

Uit figuur 2.8 blijkt dat het gemiddelde cijfer op het schoolexamen de afgelopen jaren behoorlijk constant is geweest. Dit ondanks de strengere exameneisen en vakinhoudelijke wijzigingen. Tussen 2008 en 2012 daalde het cijfer op het schoolexamen licht, waarna het later licht stijgt. Opvallend is dat het gemiddelde cijfer op het vwo altijd hoger is dan op het vmbo en havo.

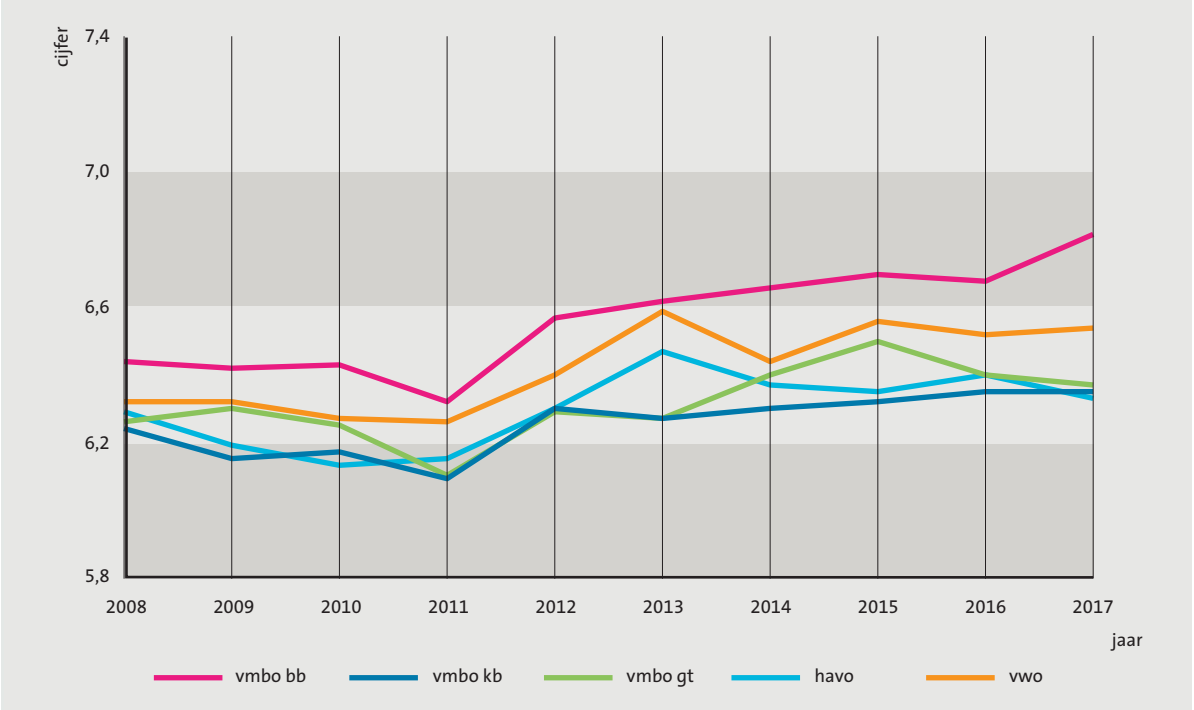
In de Examenmonitor (DUO) worden de schoolexamencijfers voor de kernvakken niet gepubliceerd. Als we kijken in het basisregister onderwijs van DUO (BRON) zien we dat de trends in de periode 2010-2017 voor de kernvakken enigszins verschillend zijn. Bij Engels zien we dat het gemiddelde cijfer vanaf 2012 op alle niveaus licht stijgt met 0,1 tot 0,2 cijferpunt. Ook het cijfer voor het schoolexamen wiskunde (wiskunde-B in havo/vwo) stijgt vanaf 2012, maar niet in vmbo-bb en vmbo-kb. Het schoolexamencijfer voor Nederlands, daarentegen, is vanaf 2012 vrijwel gelijk (zie bijlage 1 voor de grafieken).



Figuur 2.8 Gemiddeld cijfer schoolexamen 2008-2017

Figuur 2.9 laat zien dat tussen 2008 en 2011 ook het gemiddelde cijfer op het centraal examen daalt, waarna het – na de aanscherping van de exameneis – weer stijgt. Het patroon is duidelijker, maar ook grilliger dan bij de schoolexamens. Vanaf 2011 blijft de trend in het vmbo overwegend stijgend. In havo/vwo stijgt het cijfer in 2012 (aanscherping exameneis) en 2013 (introductie kernvakkenregeling). In 2014 daalt het gemiddelde cijfer om daarna ongeveer gelijk te blijven.

In de Examenmonitor wordt geen informatie gegeven over de mate waarin de examencijfers variëren tussen leerlingen. Om een indruk te krijgen van de spreiding in de cijfers op een centraal examen is in het basisregister onderwijs van DUO (BRON) het gemiddelde examencijfer na het tweede tijdvak bepaald en de standaardeviatie in de periode 2010-2017. In deze periode is het gemiddelde van alle centrale examencijfers 6,38 en de standaarddeviatie 1,13 (N=9.658.702). Dit betekent dat driekwart van de leerlingen (75,7%) een examencijfer tussen de 5 en 7,5 heeft na het tweede tijdvak.

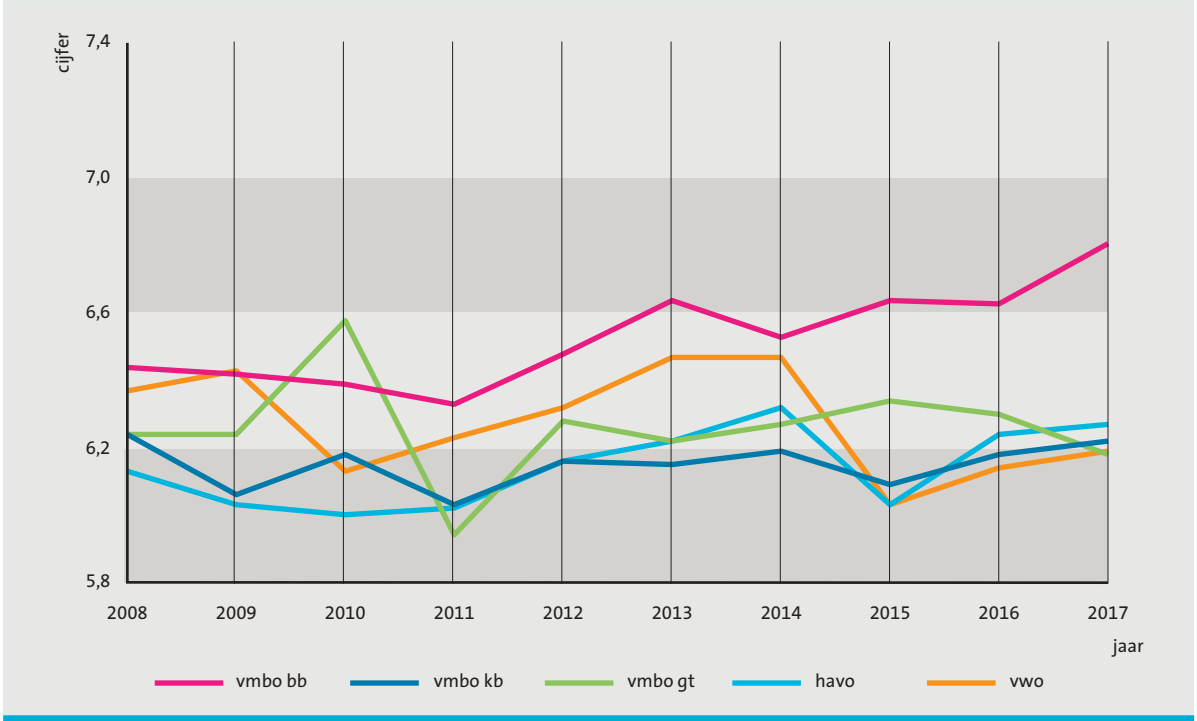


	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Rekentoets; Exameneis Nederlands aangescherpt	CE Nederlands geijkt; (Rekentoets)	(Rekentoets)	(Rekentoets)
Havo/Vwo		Heelvak/ deelvak afgeschaft (Havo)	Heelvak/ deelvak afgeschaft (Vwo)		Aange-scherpte exameneis	Kernvakken-regeling	Rekentoets	CE Neder-lands geijkt & samenvatting afgeschaft; (Rekentoets) Nieuw examenpro-gramma Natuurweten-schappen (Havo)	Rekentoets telt mee in examen-resultaat (Vwo) Nieuw examen-programma Natuurweten-schappen (Vwo)	Rekentoets telt mee in kernvakken-regel (Vwo) Nieuw examenpro-gramma Wiskunde (Havo)

Figuur 2.9 Gemiddeld cijfer centraal examen 2008-2017

Centraal examen Nederlands

In het gemiddelde cijfer voor het centraal examen Nederlands zijn tussen 2008-2017 een aantal trends zichtbaar (DUO, 2013; DUO, 2017). Zo laat figuur 2.10 tussen 2008 en 2011 een zeer wisselend beeld zien, per jaar en per schoolsoort. De periode tussen 2011 en 2014 wordt, na het aanscherpen van de exameneis, gekenschetst door een voornamelijk stijgende trend. Vervolgens blijft het gemiddelde cijfer alleen in vmbo-bb ook in de jaren na 2014 stijgen. In vmbo-kb en vmbo-gt blijft het eindcijfer na 2014 min of meer op het niveau van 2008. In havo en vwo wordt de stijgende trend in 2015 doorbroken door een duidelijke daling van het gemiddelde cijfer, om daarna weer te stijgen. DUO zegt hierover: "De ontwikkeling in het vwo kan verklaard worden uit het feit dat in 2015 de normering is afgestemd op de cesuren die horen bij de referentie-niveaus. Dit betekent dat aan de kandidaten hogere prestatie-eisen werden gesteld." (https://www.onderwijsincijfers.nl/kengetallen/voortgezet-onderwijs/deelnemersvo/eindexamens, geraadpleegd op 13-7-2018).



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Exameneis Nederlands aangescherpt	CE Neder-lands geijkt;		
Havo/ Vwo					Aange-scherpte exameneis	Kernvakken-regeling		CE Neder-lands geijkt & samenvatting afgeschaft;		

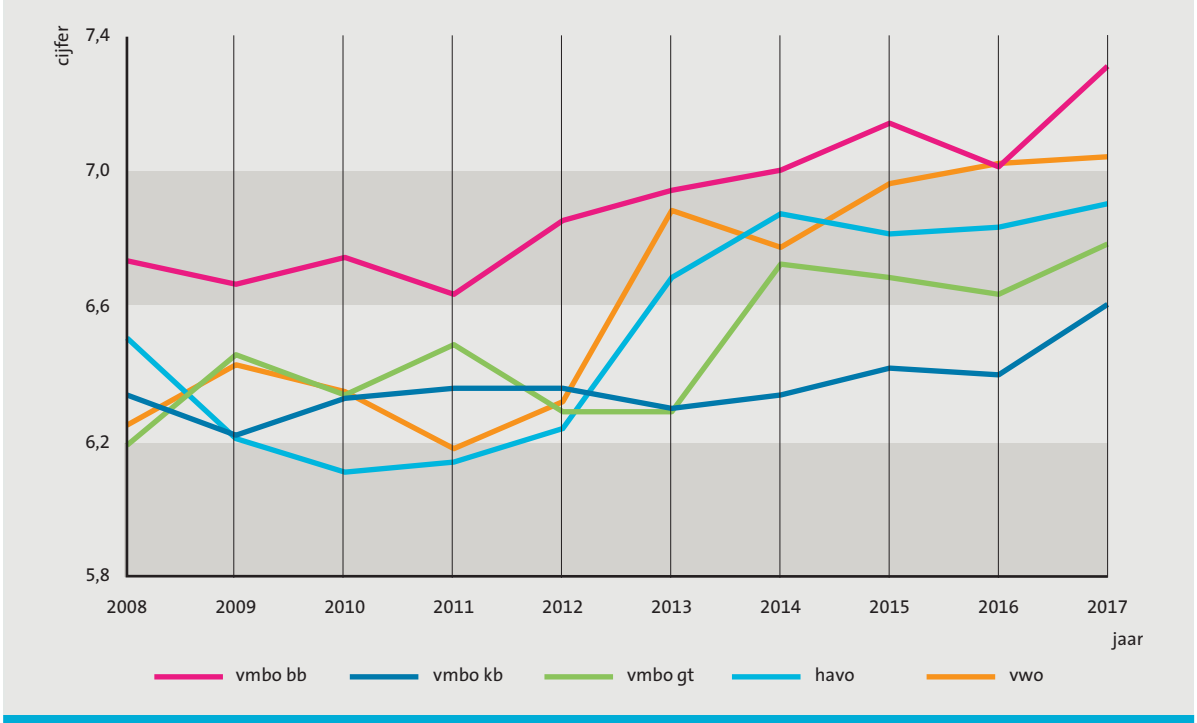
Figuur 2.10 Gemiddeld cijfer op het centraal examen Nederlands per schoolsoort.

Vergelijken we de gemiddelde cijfers in 2017 met die van 2008 (zie tabel 2.2), dan zien we in vmbo-bb een duidelijke toename met 0,37 cijferpunt. In 2008 was het gemiddelde cijfer op het centraal examen Nederlands in vmbo-bb 6,43 en in 2017 6,80. Ook in havo ligt het gemiddelde CSE-cijfer iets hoger, namelijk 0,14 cijferpunt. In vwo ligt het gemiddelde cijfer in 2017 met -0,18 cijferpunt juist lager dan in 2008. In vmbo-kb en vmbo-gt verandert het cijfer amper. Het cijfer in 2017 komt net iets onder dat in 2008 uit (respectievelijk -0,02 en -0,06 cijferpunt).

Centraal examen Engels

Het gemiddelde eindexamencijfer Engels kenmerkt zich door een beeld van stijgingen en dalingen. Focussen we eerst op vmbo-gt, dan is het beeld over de jaren heen wisselend. In 2014 is sprake van een opvallende stijging met 0,44 cijferpunt. Vmbo-kb laat tot 2016 soms een lichte daling, dan weer een lichte stijging zien. In 2017 is de stijging overigens opvallend (0,21 cijferpunt). Voor vmbo-bb is de trend vanaf 2011 duidelijk stijgend. Deze trend wordt alleen in 2016 doorbroken.

In havo en vwo is een golfbeweging zichtbaar. Het vwo laat tussen 2009 en 2011 een daling zien van de cijfers, die gevolgd wordt door een stijging. Tussen 2012 en 2013 is deze stijging geprononceerd, wat samenvalt met de introductie van de kernvakkenregeling. In havo zien we een soortgelijke golfbeweging. De daling tussen 2008 en 2010, wordt gevolgd door een stijging, die zeer duidelijk is tussen 2012 en 2013. Na 2015 is een zeer lichte stijging waarneembaar. De algehele trend over de periode 2008-2017 is echter stijgend (zie figuur 2.11): het gemiddelde cijfer op het centraal schriftelijk eindexamen is in 2017 duidelijk hoger dan in 2008. Op vmbo-bb is het verschil 0,58 cijferpunt, op vmbo-kb 0,27 cijferpunt en op vmbo-gt 0,60 cijferpunt. Havo en vwo scoren respectievelijk 0,40 en 0,80 cijferpunt hoger. Op vwo is de stijging het meest uitgesproken. Was het cijfer op vwo gemiddeld 6,25, in 2017 was het gemiddelde cijfer 7,05.



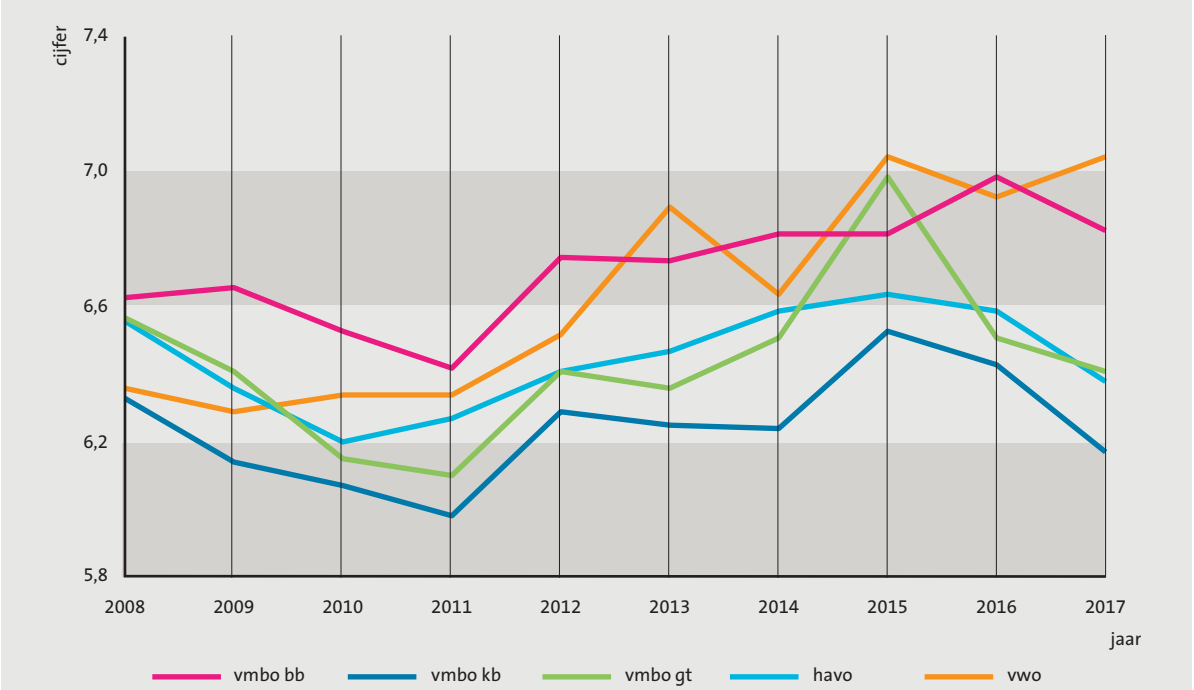
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis					
Havo/ Vwo					Aange-scherpte exameneis	Kernvakken-regeling				

Figuur 2.11 Gemiddeld cijfer op het centraal examen Engels per schoolsoort

Centraal examen wiskunde

Bij het CE-cijfer voor wiskunde zien we drie periodes. Tussen 2008 en 2011 is de trend overwegend dalend, behalve in het vwo. Na het aanscherpen van de exameneis in 2012 gaan de cijfers schommelend omhoog. In vmbo-kb, vmbo-gt en havo kentert de trend na 2015, in vmbo-bb na 2016. Voor het vwo blijft de trend overwegend stijgend. Alleen in 2014 is sprake van een duidelijke daling (-0,26 cijferpunt), in 2016 van een lichte.

Over het geheel genomen ligt het gemiddelde cijfer voor wiskunde in 2017 0,22 cijferpunt hoger dan in 2008 (zie tabel 2.2). Ook in vmbo-bb ligt het cijfer wat hoger dan in 2008 (0,20 cijferpunt), maar in havo, vmbo-gt en vmbo-kb ligt het cijfer in 2017 juist wat lager dan in 2008 (respectievelijk -0,18; -0,16;-0,16). In vwo is de verandering in het gemiddelde cijfer wiskunde meer uitgesproken. In 2008 was het gemiddeld cijfer op het centraal examen wiskunde 6,36, in 2015 was dit cijfer gestegen tot gemiddeld 7,05 (een verschil van 0,69 cijferpunt).



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Rekentoets	(Rekentoets)	(Rekentoets)	(Rekentoets)
Havo/ Vwo		Heelvak/ deelvak afgeschaft (Havo)	Heelvak/ deelvak afgeschaft (Vwo)		Aange-scherpte exameneis	Kernvakken-regeling	Rekentoets	(Rekentoets)	Rekentoets telt mee in examen-resultaat (Vwo)	Rekentoets telt mee in kernvakken-regel (Vwo) Nieuw examen-programma Wiskunde (Havo)

Figuur 2.12 Gemiddeld cijfer op het centraal examen wiskunde per schoolsoort

Tabel 2.2 Verschillen tussen de gemiddelde cijfers op het centraal schriftelijk eindexamen in 2017 en 2008

	Verschil tussen het gemiddelde cijfer in 2017 en 2008				
	School-examen	Centraal examen	Centraal examen Nederlands	Centraal examen Engels	Centraal examen wiskunde
vmbo-b	-0,08	0,38	0,37	0,58	0,20
vmbo-k	-0,06	0,11	-0,02	0,27	-0,16
vmbo-gt	-0,03	0,11	-0,06	0,60	-0,16
havo	-0,05	0,04	0,14	0,40	-0,18
vwo	-0,10	0,22	-0,18	0,80	0,69

2.3 Samenvatting en conclusie

In dit hoofdstuk hebben we de resultaten van PISA en die van de schriftelijke centrale examens bestudeerd. Hoe zien de prestaties van Nederlandse leerlingen eruit in het PISA-onderzoek en in de examens van de afgelopen tien jaar? Zijn er wijzigingen in het PISA-onderzoek en in de examenregeling die in verband gebracht kunnen worden met de gesignaleerde trends?

PISA is een internationaal vergelijkend peilingsonderzoek naar de kennis en vaardigheden van 15-jarige leerlingen. PISA is een toets met drie hoofddomeinen (leesvaardigheid, wiskundige geletterdheid en natuurwetenschappelijke geletterdheid). De toets wordt afgenomen bij een representatieve groep 15-jarige leerlingen. In elke afname ligt de nadruk op een van de drie domeinen. De toets voor het centrale domein wordt daarbij vernieuwd. De toetsing voor de niet-centrale domeinen blijft gelijk. In 2009 werd de toets voor leesvaardigheid vernieuwd, in 2012 die van wiskundige geletterdheid en in 2015 werd de toets voor natuurwetenschappelijke geletterdheid grondig herzien. In 2015 vond daarnaast de overstap plaats op een digitale afname en werden de analysetechnieken gewijzigd.

De prestaties van Nederlandse vo-scholieren op het internationale PISA-onderzoek laten een divers beeld zien. Voor leesvaardigheid blijven de prestaties op een constant niveau. Voor wiskundige geletterdheid zien we een vrij constante, lichte daling. Tussen 2012 en 2015 speelt deze overigens alleen bij havo/vwo en vmbo-bb. De prestaties bij de natuurwetenschappelijke geletterdheid, ten slotte, blijven tot en met 2012 constant en laten dan in 2015 een abrupte daling zien. Die daling geldt niet alleen voor Nederland, maar gemiddeld ook voor andere OESO-landen. Met name vmbo-leerlingen deden het dat jaar minder goed. Misschien omdat de nieuwe definitie (met de nadruk op wetenschappelijk onderzoek) minder aansluit bij hetgeen 15-jarige vmbo-leerlingen leren op school. Maar misschien ook omdat de digitale afname of nieuwe, interactieve, opgaven minder goed aansluiten bij hun digitale vaardigheden.

In 2015 werd de PISA-toetsing vrij grondig herzien, maar tot en met 2012 werkte PISA met een min of meer constante toetsinhoud, wijze van afname en techniek van analyseren en equivaleren. Bij de centrale examens waren er echter behoorlijk wat wijzigingen in de

examenregeling en het examenprogramma. Zo werden er voor vmbo-bb en kb digitale examens geïntroduceerd, werd de exameneis aangescherpt (2012) en ging de rekentoets van start (2014). Specifiek voor het eindexamen Nederlands kwam er een kernvakkenregeling (2013) en ijking aan de referentieniveaus (2015). Voor havo/vwo waren de wijzigingen nog ingrijpender. In 2007 werd 60% van het examenprogramma toegewezen aan het centraal examen en verdween de heelvak/deelvakstruktuur. Net zoals in het vmbo volgde een aanscherping van de exameneis (2012), introductie van de kernvakkenregeling (2013) en invoering van de rekentoets (2014). Het eindexamen Nederlands werd geijkt aan de referentieniveaus (2015) en de geleide samenvatting werd afgeschaft (2015). Ook werd stapsgewijs begonnen met de introductie van nieuwe examenprogramma's voor natuurwetenschappen en wiskunde.

Al deze wijzigingen en het aanscherpen van de exameneisen zien we niet sterk terug in de cijfers van het schoolexamen. Ze zijn daarentegen wel zichtbaar in de centrale examens, waar 2011 een keerpunt was. In 2012 werd de dalende trend in het gemiddelde cijfer – met de aanscherping van de exameneis – omgezet in een stijging. Niet dat die stijging zich consistent heeft doorgezet. In het vmbo vlakt de stijging af. In havo/vwo is in 2014 zelfs een dip te zien, waarna de cijfers vrijwel gelijk blijven. Ook voor de drie kernvakken is 2011 een keerpunt. Ook daar wordt in 2012 de stijging ingezet. Bij Engels blijft die stijgende lijn zich grotendeels voortzetten. Bij Nederlands zien we voor havo/vwo een dip in 2015, het jaar van de ijking aan de referentieniveaus. Bij wiskunde slaat de stijging die na 2011 is ingezet, na 2015 om in een duidelijke daling.

3 Vaardigheids-ontwikkeling en normering van de centrale examens

3 Vaardigheidsontwikkeling en normering van de centrale examens

In de motie wordt gesproken over N-termen en hoe die het zicht op vaardigheidsontwikkeling beïnvloeden. In dit hoofdstuk wordt eerst uitgelegd hoe N-termen tot stand komen. We bespreken de definitie van normeren en de geschiedenis van de normeringssystematiek. Vervolgens maken we duidelijk wat het belang is van norm-handhavingsonderzoek, waarom dat niet bij alle vakken uitgevoerd kan worden en hoe we dat oplossen. Daarna illustreren we bij enkele examenvakken de vaardigheidsontwikkeling. Deze is berekend op basis van de gegevens uit het normhandhavingsonderzoek en we vergelijken deze trends met de toegekende gemiddelde cijfers.

In Nederland willen we dat de eisen die we bij het centrale examen vo aan leerlingen stellen, van jaar tot jaar gelijk zijn. Gelijke prestaties moeten, ook over jaren heen, leiden tot een gelijk cijfer. Normhandhavingsonderzoeken leveren informatie die ervoor zorgen dat dit voldoende nauwkeurig gebeurt. Op deze manier geeft de ontwikkeling van de cijfers een goed beeld van de ontwikkeling van de vaardigheid van leerlingen. Maar hoe ontwikkelen de vaardigheden van Nederlandse leerlingen zich? Welke invloed heeft het normeringsproces op eindexamencijfers? Zorgt de Nederlandse normeringssystematiek wellicht voor een vertekening van die cijfers?

3.1 Wat is normeren?

Normeren is het toekennen van een waardering aan een score op een toets. Bekende waarderingen zijn: voldoende, matig of goed. In het Nederlandse schoolsysteem zijn waarderingen gekoppeld aan een cijfer: een 4 is onvoldoende, een 7 is ruim voldoende en een 9 is zeer goed. Het omzetten van een waardering naar een cijfer biedt de mogelijkheid om waarderingen samen te nemen. Stel, een leerling heeft twee toetsen heeft gemaakt met als oordeel een ruim voldoende (7) en een zeer goed (9). De mogelijkheid bestaat dan om het rekenkundig gemiddelde te nemen voor een eindoordeel (8).

Voor de manier waarop in Nederland scores worden omgezet naar cijfers, gelden basisregels:

1 Er is een relatie tussen het aantal scorepunten dat een leerling haalt en het cijfer dat daarbij hoort.
2 Elk scorepunt draagt evenveel bij aan het cijfer.
3 De cijferschaal loopt van 1,0 tot 10,0.
4 Cijfers moeten over de jaren heen met elkaar vergelijkbaar zijn.

Het laatste punt is een lastig punt, want resultaten zijn nooit direct met elkaar te vergelijken. Examenversies zijn namelijk nooit precies even moeilijk, en populaties van leerlingen hebben nooit precies dezelfde vaardigheid. Vergelijking is dus alleen indirect mogelijk. We doen dat na aanname van een gelijke moeilijkheid van examens, na aanname van een gelijke vaardigheid van populaties of na het uitvoeren van een normhandhavingsonderzoek.

De basisregels voor het omzetten van scores in cijfers zijn vastgelegd in de 'Regeling omzetten scores in cijfers'[9]. Deze regeling wordt jaarlijks door het College van Toetsen en Examens vastgesteld en behoeft goedkeuring van de minister.

3.2 Historische ontwikkeling van de normering

De afgelopen veertig jaar is de manier waarop onze centrale examens genormeerd worden, enkele malen bijgesteld. Deze historische ontwikkeling is van belang om te kunnen begrijpen hoe er op dit moment genormeerd wordt. Deze paragraaf beschrijft de kernpunten, een uitgebreid overzicht staat in bijlage 2.

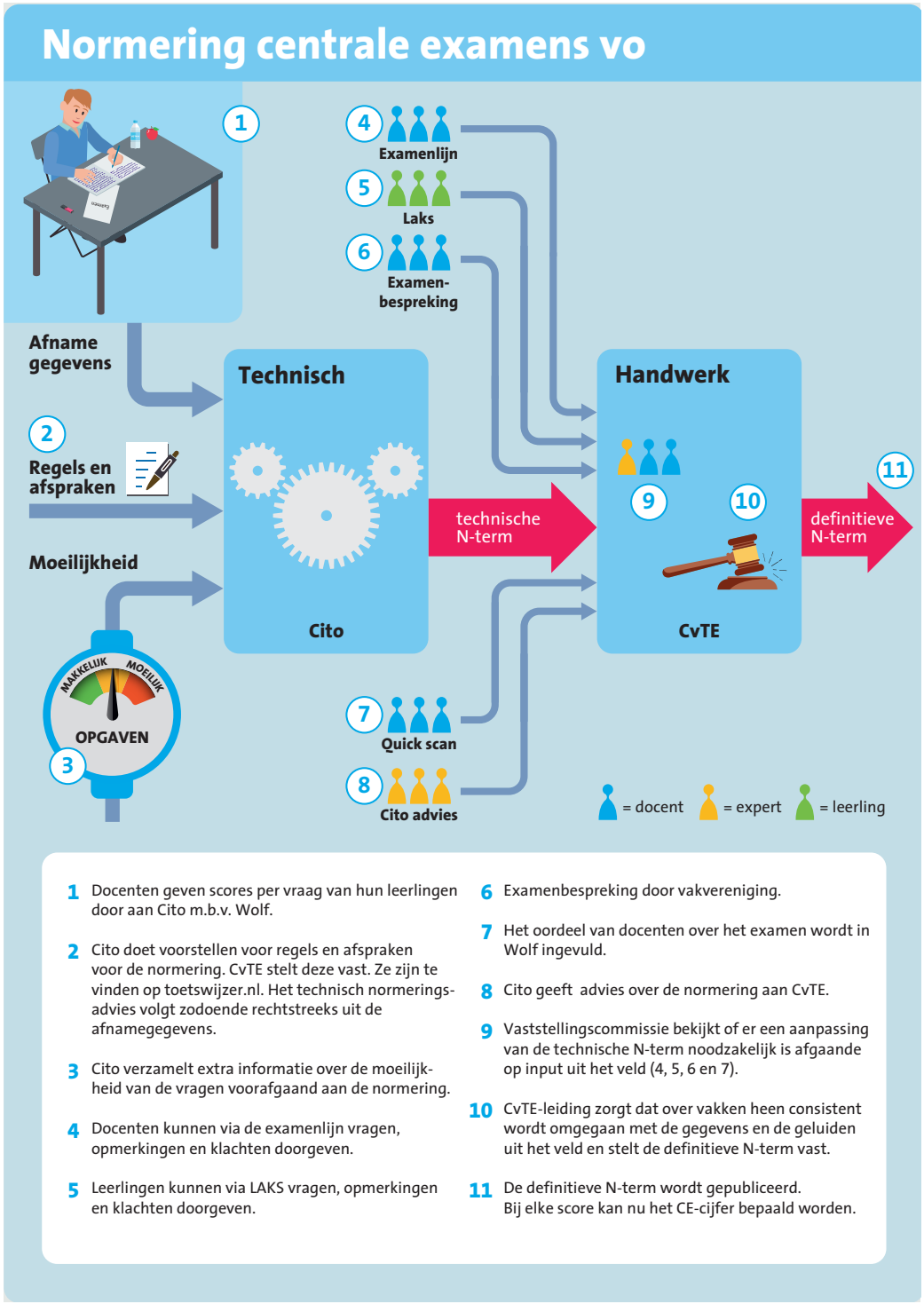
- Tot de eeuwwisseling gold in Nederland het uitgangspunt dat examens voor een bepaald vak van jaar tot jaar even moeilijk waren. Het cijfer van een leerling was afhankelijk van de moeilijkheid van het examen. Een moeilijker examen leverde gemiddeld een lager cijfer op dan een makkelijker examen.
- Rond de eeuwwisseling werden de normhandhavingstechnieken pretest en posttest geïntroduceerd. Deze normhandhavingsonderzoeken geven informatie over de moeilijkheid van een examen, waardoor voortaan genormeerd kon worden volgens het principe 'gelijke prestaties leiden tot eenzelfde cijfer'. Ook over jaren heen.
- In diezelfde tijd werd de N-term geïntroduceerd. Via dit getal tussen 0 en 2 konden verschillen in de moeilijkheid van examens worden gecompenseerd. Een behaald cijfer reflecteert op deze manier de vaardigheid van de leerling. Het cijfer werd dus onafhankelijk van de moeilijkheid van het examen: een moeilijker examen leverde hetzelfde cijfer op als een makkelijker examen.
- Voor vakken waar normhandhavingsonderzoek niet mogelijk was, gold een aanvullende aanname. Die aanname was dat de vaardigheid van de populatie niet wijzigt.
- Deze aanname is niet houdbaar wanneer de exameneisen worden aangescherpt. Daarom geldt vanaf 2012 voor vakken zonder normhandhavingsonderzoek de veronderstelling dat de vaardigheidsontwikkeling voor deze vakken vergelijkbaar is met de vaardigheids-ontwikkeling op soortgelijke vakken mét normhandhavingsonderzoek. Hierbij wordt de Fisher-methode ingezet als analysetechniek. ZIe kader op blz 40.

3.3 Het normeringsproces sinds 2012

Hoe het normeringsproces op dit moment verloopt, blijkt uit figuur 3.1. De infographic laat zien dat het normeringsproces bestaat uit twee delen: een technische analyse en een handmatige bijstelling. De technische analyse levert een technische normering op: de technische N-term. Die technische N-term is een belangrijk gegeven. Voor veruit de meeste examens wordt de technische N-term uiteindelijk namelijk ook de definitieve N-term. Voor een deel van de

9 https://wetten.overheid.nl/BWBR0037590/2019-01-01 (deze regeling wordt in de loop van voorjaar 2019 vernieuwd. Zie ook de kamerbrief van 14 januari 2019)

examens wordt de technische N-term handmatig bijgesteld in een meestal hogere of soms lagere definitieve N-term.



Figuur 3.1 Schematische weergave van het normeringsproces

3.3.1 Het normeringsproces: de technische N-term

Om de technische N-term te kunnen bepalen, werken we met input uit drie bronnen.

1 De 'Regeling omzetten scores in cijfers', zoals die jaarlijks wordt vastgesteld door het College van Toetsen en Examens. De regeling bevat de basisregels voor het omzetten van scores in cijfers.
2 De input van daadwerkelijke scores van examenkandidaten. Direct na de examens leveren docenten deze scores van hun leerlingen aan. Het is logisch dat de kwaliteit van de technische N-term mede afhankelijk is van de kwantiteit en kwaliteit van de aangeleverde data. Hoewel de aanlevering op vrijwillige basis gebeurt, werken scholen in de praktijk enthousiast mee: de scores van 85-90% van de examenleerlingen worden ingestuurd. Om de aanlevering van scores snel en soepel te laten verlopen, werkt Cito met het speciaal ontwikkelde programma Wolf.
3 De input van gegevens over de moeilijkheidsgraad van examens en examenvragen. Deze input is afkomstig uit eerder uitgevoerde normhandhavingsonderzoeken.

3.3.2 Normhandhavingsonderzoeken: vaardigheid bepalen

In de huidige normeringssystematiek vormen onze normhandhavingsonderzoeken een belangrijke gegevensbron. Via normhandhavingsonderzoeken schatten we voorafgaand aan de examens de moeilijkheid van vragen in de centrale examens in. Dat is een deel van de puzzel om de vaardigheid van de kandidaten te bepalen.

Op dit moment hanteren we drie typen normhandhavingsonderzoeken:
- Anchor in Package (AIP): deze methode wordt ingezet bij de digitale examens vmbo bb en kb. Door een deel van de opgaven geheim te houden en enkele jaren later weer in te zetten, krijgen we informatie over de vaardigheidsontwikkeling van de populatie.
- Pretest en posttest: volgens deze methode worden examenopgaven – samen met anker-opgaven – buiten de examens om voorgelegd aan leerlingen. Het levert informatie op over de moeilijkheid van examenvragen en dus de vaardigheid van de populatie.
- Standaardbepaling: hierbij schatten experts (docenten) de moeilijkheid van de vragen in. Ook daarmee kan geschat worden hoe vaardig de populatie is.

De drie schattingsmethoden zijn niet even nauwkeurig (zie tabel 3.1). Grofweg gezegd, is AIP drie keer zo nauwkeurig als een pre- of posttest, en zijn pre- en posttesten weer drie keer zo nauwkeurig als een standaardbepaling. Helaas kunnen we AIP slechts beperkt inzetten, en zijn ook pre- en posttesten niet altijd mogelijk. Pretesten werken niet bij regelmatige wisselingen van onderwerpen in het examenprogramma, of bij de introductie van een nieuw examen-programma. Posttesten werken niet bij vakken met een aanzienlijk deel handmatige correctie. Daarvoor is niet voldoende tijd. Daarom zijn we voor een aantal vakken aangewezen op de standaardbepaling. Om een zo goed mogelijk beeld te krijgen van de generieke vaardigheids-stijging, worden vanaf 2012 extra standaardbepalingen uitgevoerd.

Tabel 3.1 Standard error van de schatting van de vaardigheid van de populatie bij verschillende normhandhavingstechnieken

Normhandhavingsmethode	Standard error	Range 90% betrouwbaarheidsinterval in cijferpunten
Anchor in Package (AIP)	< 0,05	0,1
Pretest en posttest	0,10 – 0,15	0,3 – 0,4
Standaardbepaling	0,25 – 0,45	0,8 – 1,4

Uit deze tabel komt naar voren dat AIP een zeer nauwkeurige vorm van normhandhaving is. Deze techniek heeft naast dit grote voordeel ook een nadeel. Een deel van de opgaven moet namelijk geheim gehouden worden. Het centrale examen is een high-stakes toets: er hangt veel van af voor de leerling. Dit grote belang voor de leerling staat op gespannen voet met de geheimhouding. Docenten willen in alle openheid over de vragen en de correctie daarvan kunnen communiceren. Men wil het bewijs kunnen zien dat het een goed examen was.
Bij een pretest worden examenopgaven ruim (meer dan twee jaar) voor de officiële afname samen met ankeropgaven voorgelegd aan leerlingen. Het mag niet duidelijk worden waar en wanneer deze examenopgaven aan leerlingen worden voorgelegd. De opgaven moeten immers geheim blijven. Bij posttesten speelt nagenoeg geen geheimhoudingsissue. Alleen de anker-opgaven moeten geheim blijven. De opzet in pre- en posttesten is te zien in bijlage 3.
In de loop van de jaren is een patroon ontstaan welke normhandhavingstechnieken voor welke vakken worden ingezet. Het budget, het aantal leerlingen dat een examen maakt, de aard van het vak (aandeel gesloten vragen), de afnamevorm en de stabiliteit van de inhoud bepalen de keuze van de techniek. In grote lijnen worden de exacte vakken met een pretest genormeerd. De moderne vreemde talen (met uitzondering van Frans vwo) worden genormeerd met een posttest.

De meeste vakken met veel leerlingen worden op deze manier nauwkeurig van een norm voorzien. Er blijven vakken met veel leerlingen die op dit moment geen normhandhavings-onderzoek kennen. Zo blijft het een uitdaging om de vaardigheidsontwikkeling bij Nederlands, economie, geschiedenis en aardrijkskunde nauwkeurig vast te stellen. Dit zijn vakken met veel leerlingen. Vanuit het oogpunt om de beschikbare middelen zo efficiënt mogelijk in te zetten krijgen vakken met iets minder leerlingen zoals M&O, kunst, filosofie, Latijn en Spaans niet de hoogste prioriteit. Deze vakken worden hier wel genoemd als mogelijke kandidaten voor normhandhavingsonderzoek maar bedacht moet worden dat het uitvoeren van dit soort onderzoeken bij vakken met sterk wisselende inhouden heel ingewikkeld en inefficiënt wordt. Met een groter budget zou de normering verder verbeterd kunnen worden. Zo zou bij meer vakken een post- of pretest gehouden kunnen worden, de aantallen leerlingen in post- en pretest zouden vergroot kunnen worden (waarmee de nauwkeurigheid verder toeneemt) en het aantal standaardbepalingen zou vergroot kunnen worden. Ten slotte zou onderzoek kunnen leiden tot innovaties van de normering met andere nieuwe mogelijkheden (zie ook hoofdstuk 7.4.)

Tabel 3.2 laat zien welke vakken op welke manier in 2015 werden genormeerd. De tabel is bedoeld als voorbeeld, de situatie kan elk jaar iets anders zijn.

Tabel 3.2 Aanvullende activiteiten voor de normering van de centrale examens (2015)

	vak	Vmbo BB-CBT	Vmbo KB CBT	Vmbo KB CSE	Vmbo GL/GT	Havo	Vwo
1	Nederlands	AiP	sst	po	po	sst	sst
2	Engels	AiP	sst	afst GT	po	po	po
3	Wiskunde (B)	AiP	sst	sst	pr	pr	sst/pr
4	Wiskunde A					pr	pr
5	Wiskunde C						sst
6	Duits	AiP	sst	afst GT	po	po	po
7	biologie	AiP	sst	afst GT	sst	pr	sst/pr
8	economie	AiP	sst	afst GT	pr		sst
9	aardrijkskunde					sst	sst
10	M&O					sst	sst
11	Frans			afst GT	po	po	pr
12	Geschiedenis						
13	natuurkunde/nask 1	AiP		afst GT		pr	pr
14	scheikunde/nask2					pr	pr
15	kunst/beeldend				sst	sst	sst
16	Handel & adm	sst	sst				
17	Bouwtechn tim	sst	sst				
18	Consumpt bakken	sst	sst				
19	Electrotechniek	sst	sst				
20	Metalectro	sst	sst				
21	Transport & logistiek	sst	sst				
22	Verzorging	sst	sst				
23	Zorg & welzijn	sst	sst				
24	landbouw breed	sst	sst		sst		
25	dierhouderij	sst	sst				
26	ICT-route/techn/int.				sst		

CBT = Digitale examens (computer based testing), CSE = centraal schriftelijk examen, AiP = Anchor in Package, po = posttest, pr = pretest, afst GT = afstand in vaardigheid van de populaties in vmbo kb en gt, sst = standaardbepaling.

Fisher-methode

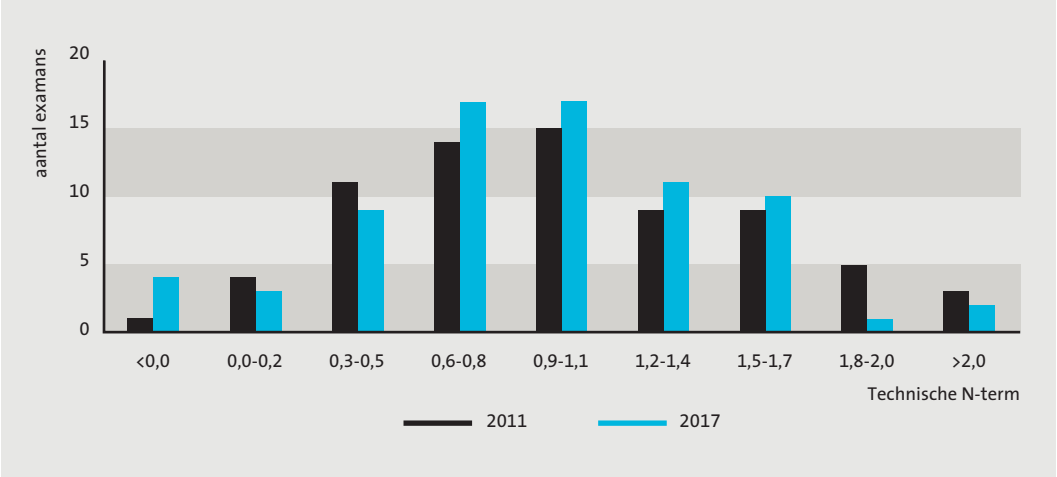
De Fisher-methode is een belangrijke analysetechniek binnen de gebruikte wijze van normhandhaving. De methode wordt gebruikt om de vaardigheidsontwikkeling van een groep leerlingen te berekenen over vakken heen. Via Fisher kunnen we de meest waarschijnlijke, generieke vaardigheidsstijging berekenen van examenkandidaten op een groep van soortgelijke vakken. De berekende vaardigheidsontwikkeling geldt vervolgens bij alle vakken van die groep, dus ook de vakken zonder normhandhavingsgegevens.
Bij de berekening wordt rekening gehouden met de nauwkeurigheid van de schatting van de vaardigheidsstijging voor elk afzonderlijk vak (zie voor meer informatie Examenblad[10]).

De Fisher-methode werkt als volgt:
- Eerst wordt een groep van coherente vakken gedefinieerd, bijvoorbeeld alle havo-niet-kernvakken.
- De normhandhavingsonderzoeken voor deze vakken worden in kaart gebracht. Voor havo gaat het om pretesten, posttesten en standaardbepalingen.
- De uitkomsten van deze onderzoeken worden samengenomen (gewogen gemiddelde).
- Dit leidt tot de meest waarschijnlijke vaardigheidsstijging van de examenkandidaten bij deze groep van vakken.
- Deze generieke vaardigheidsstijging geldt vervolgens voor alle vakken in de groep, ook de vakken zonder normhandhavingsgegevens.
- Op pagina 47 zijn de uitkomsten van de Fisher-berekeningen te zien.

3.3.3 Verdeling van de technische N-term

De moeilijkheidsgraad van de centrale examens corrigeren we in Nederland met de N-term. We zien een stijging in cijfers over de jaren heen. De vraag is of de N-termen de reële vaardigheidsontwikkeling maskeren, en de cijfers kunstmatig opgehoogd hebben. In figuur 3.2 tonen we de verdeling van de technische N-termen in 2011 en 2017. We kiezen voor 2011 omdat dit het laatste jaar was voor de introductie van de aangescherpte exameneisen. 2017 is het meest recente jaar dat we in dit rapport hebben meegenomen. De set examens is ingeperkt tot alle papieren examens van de algemene vakken van het eerste tijdvak. Deze set is qua omvang goed vergelijkbaar over de jaren heen (respectievelijk 71 en 74 examens).

10 https://www.examenblad.nl/document/de-normeringssystematiek-van-de-ce/2017/f=/De_normeringssystematiek_van_de_CEs_vo_30_november_2016.pdf.



Figuur 3.2 Verdeling van de N-term in 2011 en 2017 voor alle papieren examens van de algemene vakken

Bij een passend examen ligt de grens tussen een voldoende en onvoldoende bij ongeveer 50% van de scorepunten. Dit komt overeen met een N-term van 1,0 (voor nadere uitleg, zie bijlage 1). In de figuur zien we dat de meeste examens een technische N-term tussen 0,6 en 1,1 hebben. We zien ook dat de verdelingen van 2011 en 2017 ongeveer dezelfde vorm hebben.
De gemiddelde technische N-term in 2011 is 1,02 en de standaardafwijking 0,58. Voor 2017 is dit respectievelijk 0,93 en 0,54. Het gemiddelde is dus ongeveer gelijk gebleven, zelfs iets gedaald. Ook de spreiding is ongeveer gelijk gebleven.

Uit de formule voor de berekening van het cijfer volgt dat een gelijke N-term leidt tot een gelijk gemiddeld cijfer. Dit betekent dat de stijging van het gemiddeld cijfer op het centraal examen (zie figuur 2.9) niet verklaard kan worden met de ontwikkeling van de technische N-termen. De technische N-termen hebben de cijfers dus niet kunstmatig opgehoogd. De hogere cijfers op examens met gelijke N-termen moeten verklaard worden door andere factoren, bijvoorbeeld een hogere vaardigheid.

3.3.4 Bijstelling van de technische N-term

In de praktijk is de technische N-term de belangrijkste bepalende factor voor de definitieve N-term. Bij veruit de meeste examens wordt de technische N-term ook de definitieve N-term. Voor een deel van de examens wordt de technische N-term om inhoudelijke redenen bijgesteld. CvTE beoordeelt de inhoudelijke opmerkingen en past de normering, indien nodig, aan (zie infographic op blz 36).

Examens met bijgestelde N-term
Hoeveel examens kregen in de periode 2008 t/m 2018 een bijgestelde N-term? We hebben het onderzocht voor 2137, zowel papieren als digitale examens. Wat blijkt is dat in 413 gevallen is afgeweken van de technische N-term. Dat betekent dat voor 81% van examens de technische N-term ook de definitieve N-term werd, maar dat voor 19% een bijgestelde N-term gold.

Examens met een negatieve technische N-term
Voor 93 examens (4%) was de technische N-term negatief. Conform gemaakte afspraken, wordt zo'n negatieve technische N-term echter automatisch bijgesteld naar 0,0. In de vorige eeuw was (impliciet) de N-term altijd 1,0. Na de eeuwwisseling is hieromheen symmetrisch een interval gekozen waar de N-term binnen moet vallen. De ondergrens 0,0 is daarbij zodanig gekozen dat een leerling die 61% van de punten van het examen heeft gescoord, altijd een voldoende heeft.

Een examen met een negatieve N-term is erg makkelijk. Omdat de N-term wordt opgehoogd tot 0,0 krijgen de leerlingen in feite een hoger cijfer dan op basis van de techniek geadviseerd zou zijn. Voor de 93 examens met een negatieve N-term was de bijstelling gemiddeld +0,3.

De hoogte van de bijstelling

Vanwege hun bijzondere positie nemen we de examens met een negatieve technische N-term niet verder mee in het hier volgende deelonderzoek. In totaal blijven er in de periode 2008 t/m 2018 daarmee 2044 examens over (96%) met een niet-negatieve technische N-term. Bij 320 daarvan (16%) werd de technische N-term bijgesteld. De gemiddelde bijstelling bij deze 320 vakken is +0,09. De gegevens lieten zien dat er in de loop van de jaren geen noemenswaardige verschuivingen hebben plaats gevonden in het patroon en de hoogte van de bijstelling.

Tabel 3.3 laat zien hoe de bijstelling van de technische N-term de afgelopen jaren uitpakte. Voor ruim de helft van de bijgestelde examens gold een ophoging van 0,1. De groepen examens met een negatieve bijstelling of een positieve bijstelling van meer dan 0,1 waren ongeveer even groot. Over alle vakken gemiddeld gezien werd de technische N-term opgehoogd met 0,01. Het effect van de bijstelling op de hoogte van het gemiddeld cijfer is dus erg klein.

Tabel 3.3 Aantal vakken met bijgestelde technische N-term (2008 – 2018)

Bijstelling van technische N-term	Aantal examens	Percentage examens	Gemiddelde bijstelling
naar beneden	60	3	- 0,22
geen bijstelling	1724	84	0
ophoging van 0,1	186	9	0,10
ophoging van meer dan 0,1	74	4	0,30
Totaal	**2044**	**100**	**0,01**

Examens en vakken met bijstelling

Interessant is de vraag of bijstelling van de technische N-term op bepaalde onderwijsniveaus vaker voorkomt. Wat blijkt (zie tabel 3.4) is dat het percentage bijstellingen stijgt naarmate het onderwijsniveau hoger wordt. Op havo/vwo-niveau is het aantal examens waar wordt afgeweken van de technische N-term het hoogst (respectievelijk 33% en 25% van de bijgestelde examens).

Tabel 3.4 Examens met bijgestelde technische N-term naar onderwijsniveau

	Aantal examens	Aantal met bijstelling	Percentage	Gemiddelde bijstelling
vmbo bb	599	60	10%	0,02
vmbo kb	699	85	12%	0,12
vmbo gl/tl	257	33	13%	0,10
havo	226	75	33%	0,07
vwo	263	67	25%	0,12

Of er ook verschillen zijn tussen vakken, blijkt uit tabel 3.5. Die geeft aan dat de technische N-term veruit het vaakst wordt bijgesteld (29%) voor de exacte vakken.

Tabel 3.5 Verdeling van de vakken met een bijstelling over de verschillende groepen van vakken.

	Aantal examens	Aantal met bestelling	Percentage	Gemiddelde bijstelling
cspe	764	54	7%	0,09
exact	405	118	29%	0,10
kunst, cultuur, mens en mij	496	84	17%	0,10
talen	379	64	17%	0,05

Redenen van bijstelling

Geen enkele leerling mag de dupe worden van een inhoudelijk probleem in het examen. Duidelijk is inmiddels dat examens met een bijgestelde N-term relatief vaak voorkomen op havo/vwo en bij de exacte vakken. Je zou hieruit kunnen afleiden dat docenten bij deze vakken het meest kritisch zijn omdat de meeste inhoudelijke bijstellingen voortkomen vanuit kritiek op een vraag.

In ongeveer twee derde van de bijstellingsgevallen betreft de bijstelling een compensatie voor een onvolkomen vraag. Het gaat hier dus om ruim 200 onvolkomen vragen uit 2044 examens met naar schatting 80.000 vragen. Klachten over vermeend ondeugdelijke vragen kunnen door docenten worden doorgegeven via de examenlijn. In de bijlage is een korte notitie opgenomen met daarin een beschrijving van de procedures hieromtrent en met enkele voorbeelden van onvolkomen vragen (zie bijlage 4). Hieruit komt naar voren dat discussies over onvolkomen vragen vaak heel subtiel zijn. Er is geen sprake van goed of fout. De bijstelling voor een onvolkomen vraag wordt vaak toegepast wanneer er veel discussie is over een vraag en waarbij het vermoeden bestaat dat leerlingen op relatief grote schaal ongelijk beoordeeld zijn.
De overige redenen voor bijstelling zijn zeer divers en komen slechts sporadisch voor.

Een leerling heeft gemiddeld zeven centrale examenvakken. Bij 16% van de vakken vindt een bijstelling plaats. De gemiddelde bijstelling bij die vakken is +0,09. De gemiddelde verwachting is dat een leerling dus bij één examen door toedoen van de inhoudelijke bijstelling van de technische N-term 0,1 cijferpunt hoger krijgt.

3.4 Het technisch normeringsadvies

Voor vakken zonder normhandhavingsgegevens verloopt de bepaling van het technisch normeringsadvies anders dan voor vakken met normhandhavingsgegevens. De procedure voor de normering in het tweede tijdvak is weer anders.

3.4.1 Technisch normeringsadvies vakken zonder normhandhavingsgegevens

Sinds 2012 gaat de normering bij vakken zonder normhandhavingsgegevens in twee stappen. In de eerste stap wordt aangenomen dat de huidige populatie even vaardig is als de groep die ten grondslag ligt aan de referentiegegevens. Tot en met 2018 was dat de populatie die het referentie-examen had gemaakt. Op basis van deze aanname bepalen we voor deze vakken de voorlopige technische N-term. Na deze eerste stap werd gecontroleerd of de aanname dat beide groepen even vaardig waren, wel klopte. We onderzochten daarom of een eventuele vaardigheidsstijging of -daling had plaatsgevonden en verwerkten de uitkomst van dat onderzoek in de voorlopige technische N-term. Dit leverde de definitieve technische N-term op.

Direct na de eindexamens geven docenten via Wolf de scores van hun leerlingen aan Cito door. Dit levert een frequentieverdeling van scores op. Op basis daarvan maken we een tabel voor de verschillende waarden van de N-term en berekenen we welk cijfer een leerling bij die N-term behaald zou hebben. Zie tabel 3.6 voor een voorbeeld.

Tabel 3.6 Normeringstabel voor vwo economie (2018)

N-term	% onvoldoende	Gemiddeld cijfer	Cesuur onv / vol	Cijfer cesuur onvoldoende	Cijfer cesuur voldoende
0,5	34,1	6,0	30 / 31	5,4	5,6
0,6	29,1	6,1	29 / 30	5,3	5,5
0,7	29,1	6,2	29 / 30	5,4	5,6
0,8	24,8	6,3	28 / 29	5,4	5,5
0,9	21,0	6,4	27 / 28	5,3	5,5
1,0	21,0	6,5	27 / 28	5,4	5,6
1,1	17,0	6,6	26 / 27	5,4	5,5

Uit de tabel blijkt dat niet elke ophoging van de N-term gepaard gaat met een afname van het percentage onvoldoendes. Dat wordt veroorzaakt door de cesuurscore (kolom 4). De cesuurscore is de score waarbij het cijfer van onvoldoende naar voldoende gaat. De voorlaatste kolom geeft

aan wat het cijfer is bij de onderkant van de cesuur (net onvoldoende), de laatste kolom toont de bovenkant (net voldoende). Een voorbeeld.

- Het examen Economie vwo kende een maximumscore van 55.
- In de frequentieverdeling zien we dat 21,0% van de leerlingen 27 punten of minder heeft gescoord.
- Het cijfer bij een score van 27 en een N-term van 0,9 wordt: Cijfer = 27/55 * 9 + 0,9 = 5,3.
- Een verhoging van de N-term van 0,9 naar 1,0 levert wel een gemiddeld 0,1 hoger cijfer op, maar het cijfer bij de score 27 is nog steeds onvoldoende: 5,4. De cesuur verschuift dus niet.
- Op deze manier wordt duidelijk dat de 21% leerlingen met een score van 27 of minder nog steeds een onvoldoende krijgt, ook al stijgt de N-term van 0,9 naar 1,0.

Welke N-term adviseren we nu voor economie vwo in 2018?

- Volgens de referentiegegevens voor vwo economie scoort 24,1% van de leerlingen een onvoldoende en is het gemiddelde cijfer 6,3. Dit zijn waarden die de 2011-populatie scoorde op het referentie-examen.
- N = 0,8 komt hier het dichtst bij in de buurt. We beschouwen dit als voorlopige technische N-term.
- Uit de Fisher-methode voor de niet-kernvakken op het vwo blijkt dat leerlingen in 2018 0,1 vaardiger zijn dan de 2011-populatie. Dat betekent dat ze gemiddeld 0,1 hoger scoren op hun examen, omdat ze vaardiger zijn (en niet omdat het examen makkelijker is).
- Om recht te doen aan die grotere vaardigheid, zeggen we dat we de moeilijkheid van het examen hebben onderschat. We stellen de N-term naar boven bij. In dit geval adviseren we een definitieve technische N-term van 0,9.
- Omdat bij beschouwing van het gehele examen er geen aanleiding blijkt voor een handmatige bijstelling, wordt 0,9 de uiteindelijk vastgestelde definitieve N-term.

3.4.2 Normeringsadvies vakken met normhandhavingsgegevens

De normering van een vak met normhandhavingsgegevens baseren we op de uitkomsten van normhandhavingsonderzoeken. Het CvTE beschreef in Euclides (het vakblad van de Nederlandse vereniging van wiskundeleraren) zeer uitgebreid de normering van het vak wiskunde, een vak met een pretest: https://www.examenblad.nl/nieuws/20170627/artikel-de-n-term-in-werking/2019.
Het artikel geeft aan dat we de uitkomsten van pre- en posttesten niet klakkeloos volgen.
Een grote uitschieter wordt daarbij slechts ten dele gevolgd, omdat de uitschieter mogelijk deels is toe te schrijven aan de onzekerheden in de metingen en aan de techniek.

In de loop van de jaren is duidelijk geworden dat we de uitkomsten van pre- en posttesten met enige prudentie moeten behandelen.

- Pre- en posttesten leveren een schatting op van de moeilijkheid van de examenvragen en van vaardigheid van examenkandidaten. De N-term die hier uit voortvloeit weerspiegelt de moeilijkheid van het examen en geeft een gemiddeld cijfer dat past bij de vaardigheid van de populatie.
- Vervolgens berekenen we hoe groot het betrouwbaarheidsinterval is rond de (punt)schatting van deze N-term. Daarmee stellen we de N-term ‘conservatief’ vast. Dat betekent dat we de technische N-term kiezen die past bij de aanname van even vaardige populaties (de huidige populatie is even vaardig als de referentiepopulatie). We doen dat alleen als deze in het betrouwbaarheidsinterval ligt. Als de N-term op basis van gelijke populaties niet in het betrouwbaarheidsinterval ligt, kiezen we de technische N-term zodanig dat deze nog wel in het betrouwbaarheidsinterval maar zo dicht mogelijk bij de N-term op basis van de aanname van gelijke populaties ligt.

- Als laatste check onderzoeken we of er ten opzichte van vorig jaar een grote sprong in vaardigheid is gemaakt. De gedachte daarbij is dat, onder normale omstandigheden, het niet waarschijnlijk is dat een populatie van het ene op het andere jaar 0,3 cijferpunt sterker of zwakker wordt.

3.4.3 Normeringsadvies tweede tijdvak

De technieken zoals we die hanteren in het tweede tijdvak, wijken af van die in het eerste tijdvak. Over de tweede tijdvaknormering schreven we in 2018 een kort artikel dat via Examenblad aan de scholen werd verzonden: https://www.examenblad.nl/document/vkpmldy4x37k-toelichting-normering/2018/f=/normering_tweede_tijdvak_2018.pdf.

In principe werken we in het tweede tijdvak als volgt:
- Uitgangspunt is dat de N-term van het tweede tijdvak gelijk is aan die van het eerste tijdvak.
- In specifieke gevallen – waarbij uit onderzoek blijkt dat het tweede tijdvak moeilijker is dan het eerste – wordt de N-term van het tweede tijdvak opgehoogd.
- Als de N-term van het eerste tijdvak pas in een laat stadium wordt opgehoogd (bijvoorbeeld door een onvolkomen vraag), geldt de ophoging niet voor het tweede tijdvak. We gaan er dan vanuit, dat een dergelijke vraag in het tweede tijdvak niet voorkomt.
- Onderzoek naar het verschil in moeilijkheid tussen het eerste en tweede tijdvak baseren we op de prestaties van leerlingen die zowel het eerste als het tweede tijdvak gemaakt hebben en die in het eerste tijdvak een onvoldoende scoorden.
- Blijkt uit dit onderzoek dat het tweede tijdvak aanmerkelijk makkelijker was, dan wordt de voorlopige N-term van het tweede tijdvak volgens afspraak niet naar beneden bijgesteld. Dit geeft kandidaten vooraf een vorm van zekerheid.
- Veel vakken kennen in het tweede tijdvak weinig deelnemers. Dit maakt dat we de statistische analyse van de scores met de nodige voorzichtigheid moeten uitvoeren. We onderzoeken op dit moment de mogelijkheden om tot een verantwoorde eigenstandige normering te komen.

3.5 Vaardigheidsontwikkeling

In de inleiding van dit hoofdstuk stond het al: ‘In Nederland willen we dat de eisen die we bij het centrale examen vo aan leerlingen stellen van jaar tot jaar gelijk zijn. Gelijke prestaties moeten, ook over jaren heen, leiden tot een gelijk cijfer. Normhandhavingsonderzoeken leveren informatie die ervoor zorgen dat dit voldoende nauwkeurig gebeurt. Op deze manier geeft de ontwikkeling van de cijfers een goed beeld van de ontwikkeling van de vaardigheid van leerlingen. Maar hoe ontwikkelen de vaardigheden van Nederlandse leerlingen zich?’ Deze paragraaf laat die ontwikkeling zien aan de hand van gegevens uit normhandhavingsonderzoeken en daadwerkelijk gehaalde (en genormeerde) gemiddelde cijfers. Daarbij belichten we alle relevante vakken voor dit rapport: uiteraard de kernvakken wiskunde en Engels, maar ook de vakken biologie en natuurkunde. Ook voor het vak economie wordt de vaardigheidsontwikkeling op vmbo weergegeven zodat, naast de exacte vakken en de talen, er ook zicht is op de ontwikkeling van een vak uit het cluster ‘mens en maatschappij’. Het kernvak Nederlands ontbreekt.
Een betrouwbare analyse is hier niet mogelijk vanwege de grote inhoudelijke wijzigingen in het examen (het onderdeel schrijven verdween stapsgewijs, de meting van leesvaardigheid veranderde en er kwam een aftrek voor taalfouten). Voor zover het wel mogelijk was om een vergelijking te maken (vmbo bb en kb) zien we geen verandering in de vaardigheid van Nederlands.

3.5.1 Vaardigheidsontwikkeling 2012 – 2018

De Fisher-methode berekent de generieke vaardigheidsontwikkeling van examenkandidaten op een groep van coherente vakken. Dit is weergegeven in tabel 3.7. Daarin worden de examenpopulaties jaar na jaar afgezet tegen de vaardigheid van de kandidaten in 2011 (en eerder).

Wat blijkt is dat er met name op havo en vwo in 2012 – 2015 grote schommelingen optraden, maar dat er sinds 2015 een nieuwe stabiele situatie is ontstaan. Ook op het vmbo is het prestatieniveau de laatste drie jaar redelijk constant.

Tabel 3.7 Ontwikkeling vaardigheden eindexamenkandidaten 2012 – 2017

Groep van vakken	Vaardigheidsverschil t.o.v. populatie ≤ 2011						
	2018	2017	2016	2015	2014	2013	2012
beroepsgerichte vakken	**Nvt**	+0,1	+0,1	+0,1	+0,1	+0,1	+0,1
algemene vakken BB	**0,4**	+0,5	+0,4	+0,4	+0,3	+0,3	+0,2
algemene vakken KB	**0,1**	+0,1	+0,1	+0,0	+0,1	+0,1	+0,2
algemene vakken GT	**0,1**	+0,2	+0,2	+0,2	+0,2	+0,2	+0,2
havo niet-kernvakken	**0,0**	+0,0	+0,1	+0,0	+0,0	+0,3	+0,1
havo-kernvakken Eng, wisB	**0,7**	+0,8	+0,6	+0,7	+0,3	+0,6	
havo kernvakken Ned, wis A	**0,0**	+0,0*	+0,1*	+0,0		+0,3	
vwo niet-kernvakken	**0,1**	+0,1	+0,1	+0,2	+0,0	+0,2	+0,1
vwo-kernvakken Eng, wisA, wisB	**0,7**	+0,8	+0,6	+0,7	+0,3	+0,6	
vwo-kernvakken Ned, wisC	**0,0**	+0,1*	+0,1*	+0,2		+0,2	

ss = standaardbepaling

aip = anchor in package

pt = pre- of posttest

Vakken met gelijke kleuren vormen in de analyse een coherente groep van vakken

3.5.2 Vaardigheidsontwikkeling per vak (2008 – 2017)

Ook per vak hebben we een vergelijking gemaakt en in een grafiek weergegeven. De grafieken op de volgende pagina's geven voor de vakken met een papieren examen zicht op de jaren 2008 t/m 2017. Voor de vakken met een digitaal examen is de periode 2011 t/m 2017 weergegeven. De zwarte lijnen zijn gemaakt op basis van gegevens van Wolf. Interne kwaliteitscontrole heeft laten zien dat de gegevens in Wolf zeer nauw aansluiten met de gegevens in BRON bij DUO. De grafieken in hoofdstuk 2 (figuur 2.9 tot en met 2.11) zijn gebaseerd op BRON, DUO. De zwarte lijnen in de hier volgende grafieken laten dus hetzelfde verloop zien als de grafieken in hoofdstuk 2.

De grafieken die nu volgen maken zichtbaar tot op welke hoogte het daadwerkelijk vastgestelde cijfer overeenstemt met de vaardigheidsmeting. Om deze twee goed met elkaar te kunnen vergelijken is de vaardigheid van de verschillende populaties gebruikt om uit te rekenen wat zij

gescoord zouden hebben op een basisexamen. We hebben meestal gekozen voor 2011 als basisexamen omdat dit ook het jaar is waar we de Fisher-uitkomsten tegen hebben afgezet. De werkwijze bij het maken van de grafieken was als volgt:

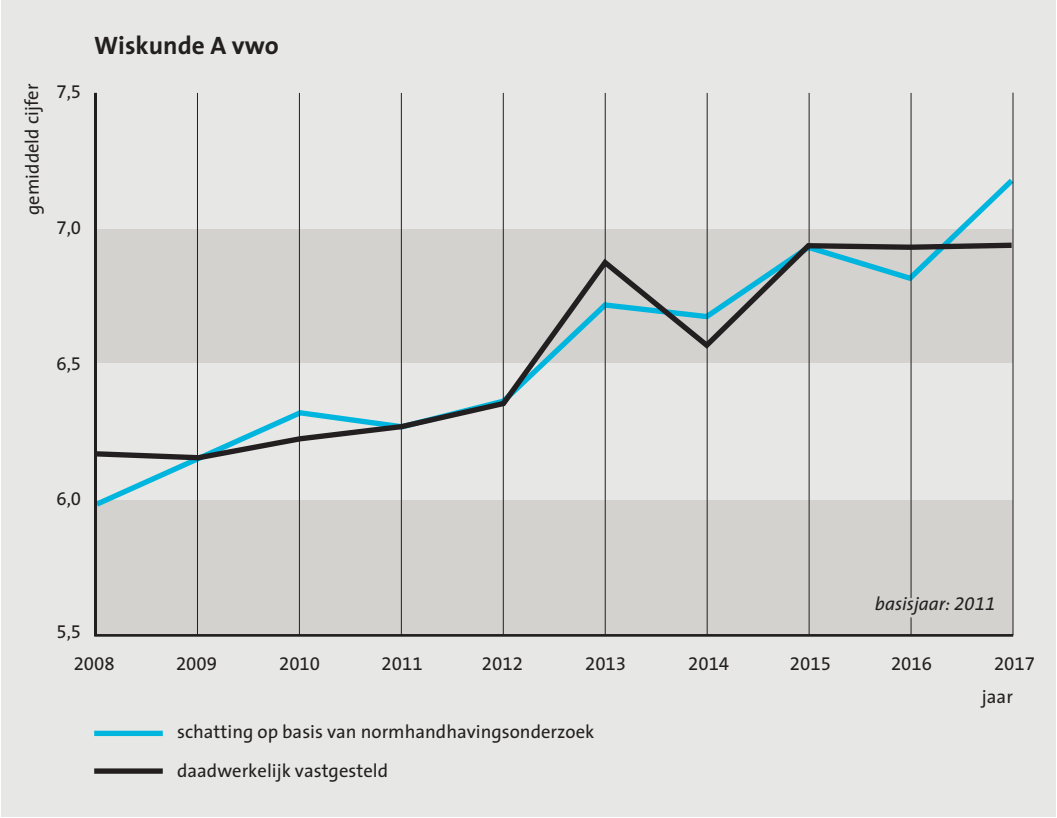
- Uitgangspunt zijn alle afnamegegevens van de ankeropgaven in de periode 2008 – 2017.
- Met behulp van de ankeropgaven zijn de relatieve moeilijkheden van alle examens in kaart gebracht. Jaren waarin de koppeling tussen het anker en het examen te zwak was, zijn weggelaten.
- Vervolgens is met behulp van de moeilijkheidsschatting van de examens, de vaardigheid van de populatie per jaar geschat.
- Vervolgens hebben we berekend welke scores een populatie, met de geschatte vaardigheid, zou halen op het basisexamen, meestal het examen uit 2011.
- Met de N-term van het basisexamen is berekend welk cijfer een populatie gemiddeld zou hebben gehaald. Deze cijfers zijn in de grafieken met een blauwe lijn weergegeven.
- Het spreekt voor zich dat het cijfer dat de 2011-populatie volgens de bovenstaande procedure zou behalen op het 2011-examen, vrijwel gelijk is aan het daadwerkelijk behaalde cijfer op het 2011-examen. Als we hadden gekozen voor 2014 als basisexamen dan kwamen de grafieken in het jaar 2014 samen.

Op basis van deze gestandaardiseerde werkwijze konden we de vaardigheden van verschillende populaties op een eerlijke manier met elkaar vergelijken. Toch, zo zal duidelijk worden, blijven er factoren die de interpretatie van de grafieken bemoeilijken. Deze factoren zijn ook van invloed op de nauwkeurigheid waarmee we de N-term kunnen bepalen, per vak.

Detailbeschouwing vwo wiskunde A

Figuur 3.2 laat zien hoe de prestatie van vwo-leerlingen in wiskunde A zich heeft ontwikkeld. De volgende zaken vallen op:

- De algehele trend is stijgend. Tussen 2012 en 2013 maakte de vaardigheid een sprong, zoals na de invoering van de kernvakkenregeling ook verwacht mocht worden. In 2014 werd echter weer een stapje terug gedaan.
- In 2016 werd de definitieve N-term 0,1 hoger vastgesteld dan de technische N-term. Reden was een onvolkomen vraag in het examen.
- De vaardigheid van de populatie in 2017 ligt ruim 0,9 cijferpunt boven die van de populatie in 2011.
- In 2017 zien we dat de vaardigheidsmeting een hoger cijfer had geadviseerd dan daadwerkelijk in de praktijk is toegekend. De reden dat die stijging van 2016 naar 2017 in de praktijk niet is gevolgd, heeft te maken met de manier waarop het gebruik van het betrouwbaarheidsinterval uitpakt (zie 3.4.2).

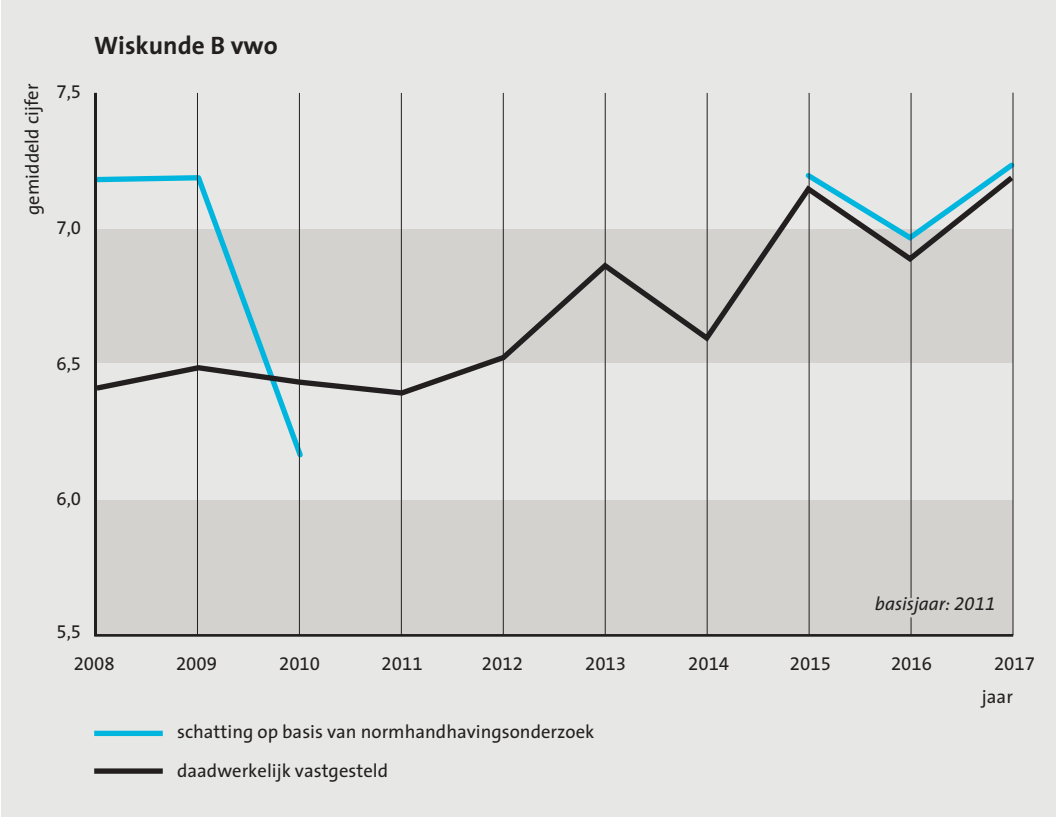


Figuur 3.2 Ontwikkeling van de prestatie van vwo examenkandidaten wiskunde A. Weergegeven in blauw is het gemiddeld cijfer dat de verschillende populaties behaald zouden hebben wanneer zij het 2011-examen zouden hebben gemaakt. Dit is berekend aan de hand van de pretestgegevens. De zwarte lijn geeft het gemiddeld cijfer weer zoals dat in de praktijk is waargenomen.

Detailbeschouwing vwo wiskunde B

Hoe vaardig vwo-leerlingen de afgelopen jaren waren in wiskunde B, blijkt uit figuur 3.3. Daarin vallen de volgende zaken op:

- Opvallend is de hoge vaardigheidsscore in 2008 en 2009, die niet tot uitdrukking komt in de vastgestelde cijfers. De verklaring is dat de scores gebaseerd zijn op leerlingen die wiskunde B1,2 deden. Deze leerlingen waren erg goed in wiskunde; het aantal studielasturen voor wiskunde B1,2 lag aanzienlijk hoger dan voor wiskunde B vanaf 2010. Bovendien ging wiskunde B1,2 vanaf 2010 min of meer samen met wiskunde B1. Daardoor was de populatie leerlingen met wiskunde B heterogener dan de populatie met wiskunde B1,2, lees zwakker.
- Voor wiskunde B1,2 gold in 2008 en 2009 een strengere norm dan voor wiskunde B vanaf 2010. Dit zorgde ervoor dat, ondanks dat de vaardigheden van de populaties vanaf 2010 een stuk lager lagen dan in 2008 en 2009, er toch een gemiddeld cijfer van 6,4 uit kwam. Gezien de geringere onderwijstijd en het samengaan met de B1-populatie leek dit redelijk.
- In 2011 t/m 2014 is het cijfer volgend uit de pretestresultaten niet weergegeven. Te weinig opgaven uit het examen en de pretest waren exact gelijk. Wiskunde B werd als kernvak vergeleken met de generieke vaardigheidsstijging voor alle kernvakken havo/vwo. Omdat voor een aantal kernvakken de vaardigheid in 2014 lager lag dan in 2013, werd de normering voor wiskunde B lager vastgesteld.
- Van 2015 t/m 2017 was het percentage gepreteste opgaven in het examen wel weer voldoende hoog. In deze jaren was de meting van de vaardigheidsontwikkeling bij wiskunde B weer leidend voor de ontwikkeling van het gemiddeld cijfer.

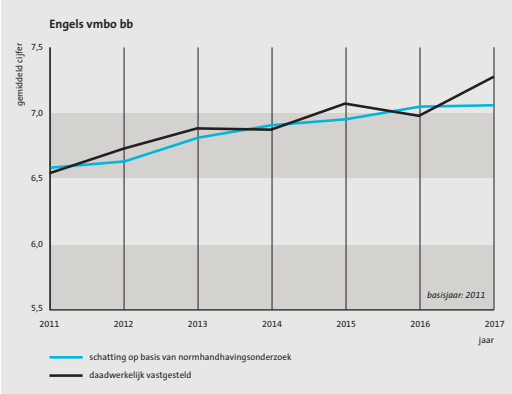


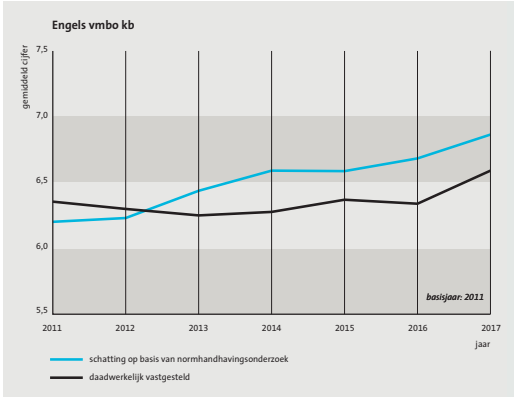
Figuur 3.3 Ontwikkeling van de prestatie van vwo examenkandidaten wiskunde B. Weergegeven in blauw is het gemiddeld cijfer dat de verschillende populaties behaald zouden hebben wanneer zij het 2011-examen zouden hebben gemaakt. Dit is berekend aan de hand van de pretestgegevens. De zwarte lijn geeft het gemiddeld cijfer weer zoals dat in de praktijk is waargenomen.

Detailbeschouwing Engels

De vaardigheidsontwikkeling voor het vak Engels blijkt uit de figuren 3.4 t/m 3.8. Daarin valt het volgende op:

- Bij Engels havo en vwo is de plotselinge stijging in 2013 duidelijk zichtbaar.
- Hoewel in vmbo gt Engels geen kernvak is, is ook hier sprake van een duidelijke stijging. De verklaring hiervoor zou kunnen liggen in de grotere aandacht voor Engels, bijvoorbeeld al in het primair onderwijs. Ook komen de leerlingen de afgelopen tien jaar vaker in aanraking met Engels in het dagelijks leven.
- Het laatste jaar (2017) lijkt de vaardigheid in Engels weer iets te dalen op havo, vwo en vmbo gl/tl.

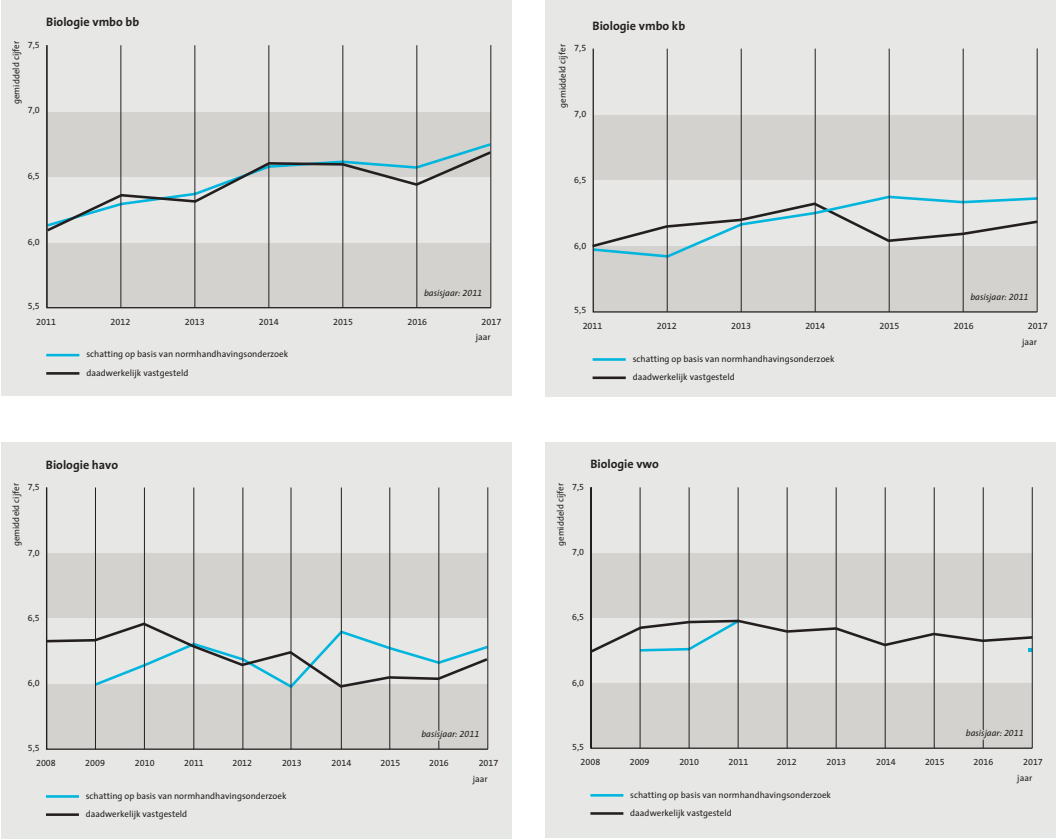




Figuren 3.4 t/m 3.8 *Ontwikkeling van de prestatie van examenkandidaten op het vak Engels. Weergegeven in blauw is het gemiddeld cijfer dat de verschillende populaties behaald zouden hebben wanneer zij het 2011-examen zouden hebben gemaakt. De zwarte lijn geeft het gemiddeld cijfer weer zoals dat in de praktijk is waargenomen.*

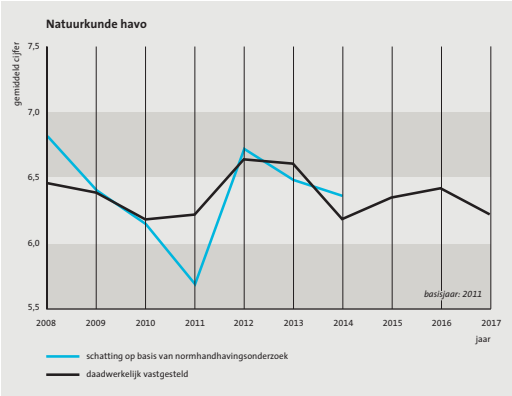
Detailbeschouwing biologie en natuurkunde

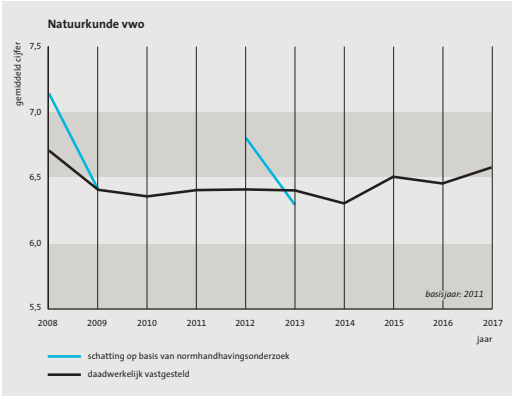
Hoe het zit met de vaardigheidsontwikkeling in biologie en natuurkunde, laten respectievelijk figuur 3.9 tot en met 3.11 en 3.12/3.13 zien. Voor biologie havo en vwo valt op dat de vaardigheid redelijk stabiel lijkt. Over de jaren heen is geen duidelijke trend zichtbaar. Bij vmbo biologie lijkt er wel een lichte stijging van de vaardigheid te zijn.



Figuren 3.9 tot en met 3.12 *Ontwikkeling van de prestatie van examenkandidaten op het vak biologie. Weergegeven in blauw is het gemiddeld cijfer dat de verschillende populaties behaald zouden hebben wanneer zij het 2011-examen zouden hebben gemaakt. De zwarte lijn geeft het gemiddeld cijfer weer zoals dat in de praktijk is waargenomen.*

Voor natuurkunde geldt, net als bij wiskunde B, dat in 2008 nog het heelvak natuurkunde 1,2 gegeven werd. Dit was een zeer vaardige populatie. Na 2010 zien we bij havo en vwo een aantal uitschieters. Bijvoorbeeld in 2011 op havo en in 2012 op vwo. Waarom de vaardigheid in deze jaren opeens zo anders was dan de jaren ervoor en erna, is niet duidelijk. Om deze reden zijn deze uitschieters niet gevolgd in de normering.





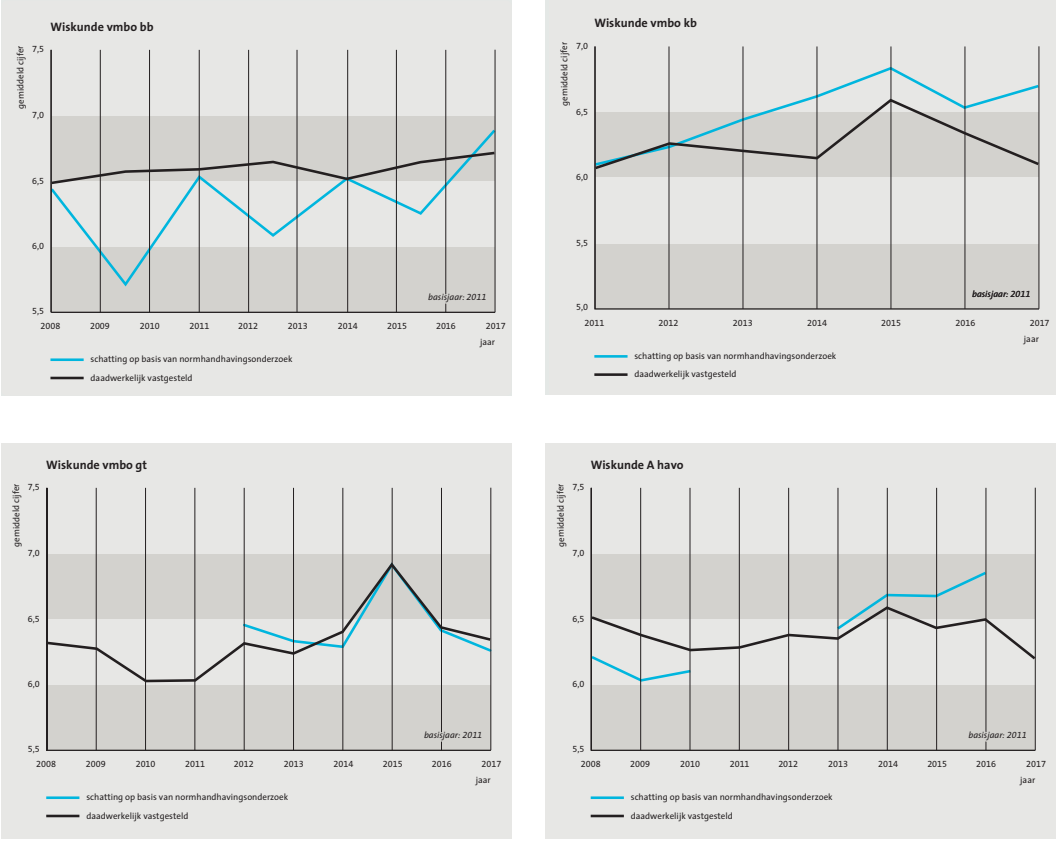
Figuren 3.13 en 3.14 *Ontwikkeling van de prestatie van examenkandidaten op het vak natuurkunde. Weergegeven in blauw is het gemiddeld cijfer dat de verschillende populaties behaald zouden hebben wanneer zij het 2009-examen zouden hebben gemaakt. De zwarte lijn geeft het gemiddeld cijfer weer zoals dat in de praktijk is waargenomen.*

Detailbeschouwing wiskunde vmbo en wiskunde A havo

De vaardigheidsontwikkeling die vmbo-leerlingen laten zien in wiskunde, is weergegeven in de figuren 3.13 t/m 3.16 (Wiskunde B havo wordt niet getoond omdat hier onvoldoende gegevens beschikbaar waren).

In de figuren valt het volgende op:

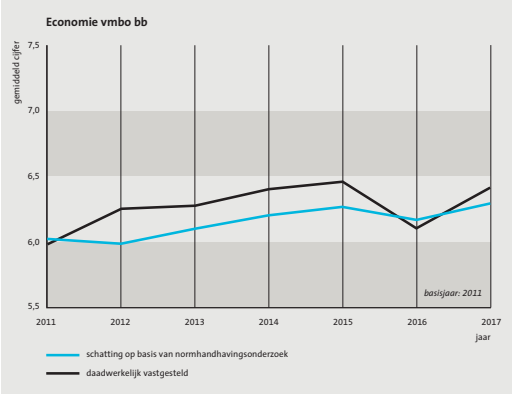
- Bij wiskunde op het vmbo bb en kb lijkt een lichte stijging zichtbaar. Bij vmbo gl/tl is een merkwaardige piek te zien in 2015. Hiervoor hebben we geen verklaring kunnen vinden.
- Wiskunde A havo is in 2009 overgestapt op het nieuwe examenprogramma. Sinds die tijd is een langzame stijging in vaardigheid zichtbaar.
- De plotselinge stijging zoals op vwo in 2013 lijkt te ontbreken.
- 2017 was het eerste jaar volgens het nieuwe cTwo-examenprogramma. Om deze reden is geen resultaat van normhandhavingsonderzoeken beschikbaar in dat jaar.

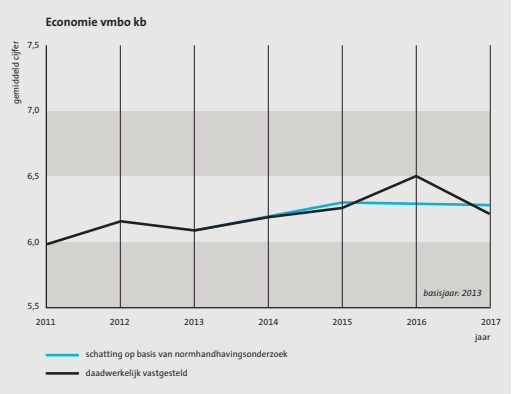


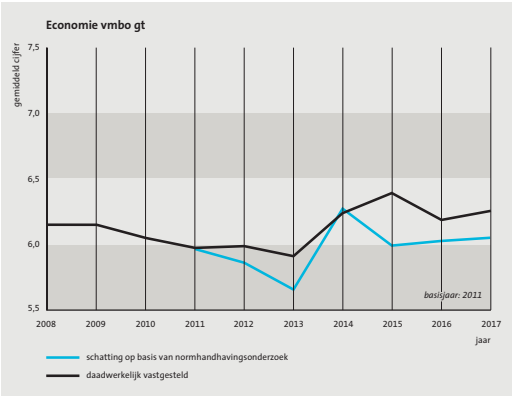
Figuren 3.15 tot en met 3.18 *Ontwikkeling van de prestatie van examenkandidaten op het vak wiskunde. Weergegeven in blauw is het gemiddeld cijfer dat de verschillende populaties behaald zouden hebben wanneer zij het 2011-examen zouden hebben gemaakt. De zwarte lijn geeft het gemiddeld cijfer weer zoals dat in de praktijk is waargenomen.*

Detailbeschouwing economie

Hoe vaardig vmbo-leerlingen de afgelopen jaren zijn geweest in economie, blijkt uit de figuren 3.17 t/m 3.19. Daarin valt op dat, net als bij biologie, er op vmbo bb en kb een lichte stijging waarneembaar is. Voor economie vmbo gl/tl is er een vreemde schommeling zichtbaar. Hier hebben we geen verklaring voor kunnen vinden.







Figuren 3.19 tot en met 3.21 *Ontwikkeling van de prestatie van examenkandidaten op het vak economie. Weergeven in blauw is het gemiddeld cijfer dat de verschillende populaties behaald zouden hebben wanneer zij het 2011- (of 2013-)examen zouden hebben gemaakt. De zwarte lijn geeft het gemiddeld cijfer weer zoals dat in de praktijk is waargenomen.*

3.6 Samenvatting en conclusie

De normering van de centrale examens in Nederland heeft zich in de loop van de tijd ontwikkeld. Sinds het begin van deze eeuw helpen normhandhavingsonderzoeken om de vaardigheids-ontwikkeling van verschillende examenpopulaties in beeld te brengen. De onderzoeken leveren relevante informatie, maar de uitkomsten bevatten nog steeds enige onzekerheid. De mate van onzekerheid is het kleinst bij de digitale examens die genormeerd worden met behulp van Anchor in Package. Veel exacte vakken hebben een pretest en veel moderne vreemde talen een posttest. Met deze technieken is het goed mogelijk om de moeilijkheid van het examen, en daarmee de N-term, te bepalen. Deze technieken kunnen echter niet bij alle vakken worden toegepast. Vakken met (regelmatig) wisselende inhouden lenen zich niet goed voor een pretest omdat de pretest twee jaar voor de daadwerkelijke afname wordt gehouden. Deze vakken lenen zich ook niet voor een posttest omdat de tijd tussen de afname van het examen en de uitslag te

kort is om tot een goede correctie te komen. Toch blijven we bij deze vakken zoeken naar goede alternatieven voor de normering. We streven ernaar om bij zo veel mogelijk vakken normhandhavingsonderzoeken te laten plaats vinden om zodoende de N-term zo nauwkeurig mogelijk vast te stellen.
De technische N-term wordt bij de vakken met normhandhavingsonderzoeken berekend aan de hand van de resultaten van die onderzoeken. Bij vakken waar dit niet gebeurt, wordt er in eerste instantie van uitgegaan dat de vaardigheid van de populatie tussen twee opeenvolgende jaren niet zal verschillen. Het percentage onvoldoende en het gemiddeld cijfer zoals die zijn gescoord op een referentie-examen, zijn leidend bij het vaststellen van de voorlopige technische N-term. Door middel van het gebruik van de Fisher-methode wordt nagegaan of er zich een generieke vaardigheidsstijging heeft voorgedaan. Deze wordt dan ook van toepassing op de vakken zonder normhandhavingsgegevens.
De definitieve N-term wijkt in een beperkt aantal gevallen af van de technische N-term. Bij één op de zes vakken wordt de N-term met gemiddeld 0,1 verhoogd. In de meeste gevallen is deze ophoging nodig omdat er inhoudelijk iets aan de hand was met een vraag.
De metingen van de vaardigheid laten zien dat de meeste vakken een stabiele of stijgende vaardigheid kennen, zoals we die ook al zagen in de ontwikkeling van de cijfers zoals die is gerapporteerd in de verschillende jaargangen van de examenmonitor van DUO, in hoofdstuk 2. Met name bij wiskunde en Engels op havo en vwo is de vaardigheid sterk gestegen sinds 2012.

4 Vaardigheden – een inhouds-analyse

4 Vaardigheden – een inhoudsanalyse

Een mogelijk verschil tussen het internationale PISA-onderzoek en de Nederlandse centrale eindexamens is gelegen in wat er wordt getoetst en hoe. Omdat PISA zinvolle internationale vergelijkingen mogelijk moet maken, meet PISA niet de beheersing van het schoolcurriculum, maar de kennis en vaardigheden die nodig zijn voor het volwassen leven (OECD, 2000). De landelijke examens worden daarentegen nauw afgestemd op het examenprogramma zoals dat is beschreven in eindtermen.

Om het verschil met curriculumgebonden kennis te benadrukken, spreekt PISA van 'wiskundige geletterdheid' en 'natuurwetenschappelijke geletterdheid'. Maar in hoeverre is die internationale toetsing van bijvoorbeeld 'natuurwetenschappelijke geletterdheid' te vergelijken met de nationale toetsing van de vakken natuurkunde, scheikunde en biologie? Daar komt bij dat PISA soms gebruikmaakt van een andere vorm van toetsing. Zo was de PISA-afname in 2015 geheel digitaal en werden er innovatieve vraagtypen gebruikt (OECD, 2015). Het is mogelijk dat ook dit verschillen oplevert in de prestaties van leerlingen.

Dit hoofdstuk bevat een vergelijkende inhoudsanalyse. We vergelijken de vaardigheidsdefinities en opgaven in PISA met die van de eindexamens. We doen dat voor leesvaardigheid, wiskundige geletterdheid en natuurwetenschappelijke geletterdheid.

4.1 Leesvaardigheid – een analyse

4.1.1 Wat is leesvaardigheid?

Leesvaardigheid in het examenprogramma Nederlandse taal
In het Examenprogramma Nederlandse taal vmbo is leesvaardigheid een van de zeven te toetsen domeinen. In tabel 4.1 worden de eindtermen van Leesvaardigheid in vmbo beschreven. Op havo/vwo is leesvaardigheid een van de zes domeinen in het examenprogramma van Nederlandse taal en literatuur (Examenprogramma Nederlandse taal en literatuur havo/vwo, p. 1). Het domein leesvaardigheid bevat drie subdomeinen: (1) analyseren en interpreteren, (2) beoordelen en (3) samenvatten (zie tabel 4.2).

Tabel 4.1 Leesvaardigheid in het Examenprogramma Nederlandse taal vmbo (vanaf 2014)

De kandidaat kan:
- leesstrategieën hanteren
- compenserende strategieën kiezen en hanteren
- functie van beeld en opmaak in een tekst herkennen
- het schrijfdoel van de auteur aangeven
- (en de talige middelen die hij hanteert om dit doel te bereiken, voor vmbo kb/gt)
- een tekst indelen in betekenisvolle eenheden en de relaties tussen die eenheden benoemen
- het hoofdonderwerp en de hoofdgedachte van een tekst aangeven
- (en een samenvatting geven, voor vmbo kb/gt)
- een oordeel geven over de tekst en dit oordeel toelichten.

Tabel 4.2 Leesvaardigheid in het Examenprogramma Nederlandse taal havo/vwo

Subdomein A1: Analyseren en interpreteren
(1) De kandidaat kan:
- vaststellen tot welke tekstsoort een tekst of tekstgedeelte behoort;
- de hoofdgedachte van een tekst(gedeelte) aangeven;
- relaties tussen delen van een tekst aangeven;
- conclusies trekken met betrekking tot intenties, opvattingen en gevoelens van de auteur;
- standpunten en soorten argumenten herkennen en onderscheiden;
- argumentatieschema's herkennen.

Subdomein A2: Beoordelen
(2) De kandidaat kan een betogende tekst of betogend tekstgedeelte op aanvaardbaarheid beoordelen en in deze tekst drogredenen herkennen.

Subdomein A3: Samenvatten
(3) De kandidaat kan teksten en tekstgedeelten beknopt samenvatten.

Leesvaardigheid in PISA

In PISA is leesvaardigheid gedefinieerd als 'het begrijpen, gebruiken, reflecteren en betrokken zijn bij geschreven teksten om je doelen te bereiken, je kennis en potentieel te verruimen, en deel te nemen aan de maatschappij'. Leesvaardigheid heeft in PISA niet alleen betrekking op het decoderen en letterlijk begrijpen van tekst, maar houdt ook het interpreteren en reflecteren op informatie in, en het vermogen om lezen te gebruiken om je doelen te kunnen bereiken. De focus in PISA is op lezen om te leren, meer dan op leren om te lezen (Cito, 2010).

In 2009 veranderde de definitie en toetsing van leesvaardigheid in PISA (OECD, 2000; OECD, 2009). De belangrijkste wijziging was de opname van het lezen van elektronische teksten (OECD, 2009), maar dit onderdeel werd in Nederland in 2009 niet afgenomen.

Definitie Leesvaardigheid in 2000/2003/2006:
De vaardigheid om schriftelijke informatie te begrijpen en weloverwogen te gebruiken, in de eerste plaats om een concreet doel te bereiken en meer in het algemeen om kennis en vaardigheden te ontwikkelen en om aan de samenleving deel te nemen.

Definitie Leesvaardigheid in 2009/2012/2015:
Het begrijpen, gebruiken, reflecteren en betrokken zijn bij geschreven teksten om je doelen te bereiken, je kennis en potentieel te verruimen, en deel te nemen aan de maatschappij.

In PISA heeft de meting van leesvaardigheid drie taakaspecten: het soort teksten (de reeks van materialen die wordt gelezen), de leesprocessen (de cognitieve benadering die bepaalt hoe de lezer met de tekst omgaat) en de situaties (de brede reeks contexten of doelen waarvoor het lezen plaats vindt) (zie tabel 4-3). Deze drie taakaspecten zorgen voor een brede dekking van de meting en door manipulatie van tekstsoort en leesprocessen wordt de moeilijkheid van de taak gevarieerd.

Tabel 4.3 Kenmerken en taakaspecten leesvaardigheid PISA 2009 (Cito, 2010, p. 29)

Definitie en belangrijkste kenmerken	• Het begrijpen van, gebruiken van, reflecteren op en betrokken zijn bij geschreven teksten om je doelen te bereiken, je kennis en potentieel te verruimen, en deel te nemen aan de maatschappij. • Naast decoderen en letterlijk begrijpen, houdt leesvaardigheid ook in interpreteren en reflecteren, en het vermogen om lezen te gebruiken om je doelen in het leven te bereiken. • De focus in PISA is op lezen om te leren, meer dan op leren om te lezen, vandaar dat de leerlingen niet beoordeeld worden op de meest basale aspecten van leesvaardigheid.
Tekstsoort	• Doorlopende teksten, waaronder verschillende soorten proza zoals vertelling, expositie en argumentatie. • Niet-doorlopende teksten, waaronder grafieken, formulieren en lijsten. • Gecombineerde teksten, een combinatie van doorlopende en niet-doorlopende teksten. • Multipele teksten, waaronder onafhankelijke teksten (in hetzelfde of in een ander format) die voor bepaalde doeleinden tegenover elkaar gezet zijn.
Leesprocessen	• Zoeken en vinden • Integreren en interpreteren • Reflecteren en evalueren • Complex, bijv. het vinden, evalueren en integreren van informatie uit verschillende soorten (elektronische) teksten
Situaties	Het gebruik waarvoor de tekst is geconstrueerd: • Persoonlijk • Publiek • Schools • Beroepsmatig

Overlap en verschillen
Er is een sterke overlap tussen ‘leesvaardigheid’ in PISA (2015) en de examenprogramma’s (vmbo, havo en vwo 2018). Naar de letter van de beschreven vaardigheden is er nauwelijks verschil. PISA (2015) stuurt echter sterk aan op reflectie en het functioneel gebruik van teksten in het maatschappelijke verkeer en voor het gebruik voor persoonlijke doelen (‘lezen om te leren’). De examenprogramma’s sturen daarentegen sterk aan op (abstracte) tekstanalyse en tekstbeschouwingen (met bijbehorende tekstanalytisch begrippenapparaat, waaronder ook het argumentatieve jargon), in het PISA-programma ook wel aangeduid met ‘leren om te lezen’.

4.1.2 Hoe wordt leesvaardigheid getoetst?

De scoringssystematiek voor leesvaardigheid is bij PISA en de centrale examens ongeveer hetzelfde: overwegend één scorepunt per correct zelfstandig antwoordelement. In de beoordelingspraktijk is PISA misschien iets coulanter. Bovendien maakt PISA bij de beoordeling van de antwoorden gebruik van een voor PISA specifiek coderingssysteem.

Qua vraagtypen zijn er geen noemenswaardige verschillen tussen PISA-vragen en examenvragen. Vrijwel alle leesvaardigheidsvragen die in PISA 2009 worden gesteld (tenminste voor zover vrijgegeven en beschikbaar gesteld), zouden deel uit kunnen maken van de centrale examens leesvaardigheid (vmbo). Uitzondering zijn vragen die een beroep doen op 'buitentekstuele' kennis, zie figuur 4.1 voor een voorbeeld. Daarin wordt gevraagd naar 'andere factoren' die niet in de tekst zelf worden genoemd, maar zijn ontleend aan een werkelijkheid 'buiten de tekst' (zoals persoonlijke ervaringen). Dit is bij examenvragen Nederlands ongebruikelijk.

VEILIGHEID VAN MOBIELE TELEFOONS

Hoofdpunt

Eind jaren 90 zijn er tegenstrijdige berichten verschenen over de gezondheidsrisico's van mobiele telefoons.

Hoofdpunt

Miljoenen euro's zijn er nu geïnvesteerd in wetenschappelijk onderzoek om de effecten van mobiele telefoons te onderzoeken.

	Zijn mobiele telefoons gevaarlijk?	
	Ja	**Nee**
1.	Radiogolven die afgegeven worden door mobiele telefoons kunnen lichaamsweefsel opwarmen, met schadelijke gevolgen.	Radiogolven zijn niet sterk genoeg om door warmte schade te veroorzaken aan het lichaam.
2.	Magnetische velden die veroorzaakt worden door mobiele telefoons kunnen de manier aantasten waarop je lichaamscellen werken.	De magnetische velden zijn ongelofelijk klein en hebben dus waarschijnlijk geen effect op de cellen in ons lichaam.
3.	Mensen die lange gesprekken voeren met mobiele telefoons klagen soms over vermoeidheid, hoofdpijn en concentratieverlies.	Deze effecten zijn in laboratoriumomstandigheden nooit waargenomen en komen misschien door andere factoren in de moderne levensstijl.
4.	Gebruikers van mobiele telefoons hebben 2,5 keer zoveel kans om kanker te krijgen in hersengebieden bij het oor dat in contact staat met het mobieltje.	Onderzoekers erkennen dat het onduidelijk is of deze toename te maken heeft met het gebruik van mobiele telefoons.
5.	Het Internationaal Bureau voor Kankeronderzoek heeft een verband gevonden tussen jeugdkanker en hoogspanningsdraden. Net als mobiele telefoons zenden hoogspanningsdraden ook straling uit.	De straling die door hoogspanningsdraden veroorzaakt wordt, is een ander soort straling, met veel meer energie dan die van mobiele telefoons afkomt.
6.	Radiofrequentiegolven die lijken op die in mobiele telefoons veranderden het genenpatroon in draadwormen.	Wormen zijn geen mensen, het is dus helemaal niet zeker dat onze hersencellen op dezelfde manier zullen reageren.

Vraag 6: VEILIGHEID VAN MOBIELE TELEFOONS R414Q06 – 019

Bekijk punt 3 in de kolom **Nee** van de tabel. Wat zou in deze context een van deze "andere factoren" kunnen zijn? Leg uit waarom je dit antwoord gegeven hebt.

.....

Figuur 4.1 Voorbeeldopgave uit PISA 2009 die een beroep doet op 'buitentekstuele' kennis (Cito, 2010, p. 166-168)

Andersom, echter, lenen veel vragen uit de centrale examens Nederlands zich niet voor opname in een PISA-toets. Vooral tekstbeschouwelijke vragen (zoals vragen naar de functie van alinea's, naar argumenten en drogredenering) zijn niet gangbaar in PISA.

Sinds de invoering van de vmbo-examens en de invoering van de Tweede Fase hebben de centrale examens Nederlands een reeks van (kleine) wijzigingen ondergaan. Dit voortdurende proces van verbetering, aanpassing en voortschrijdend inzicht voor wat betreft de vakinhoud is van alle tijden. Inmiddels bevatten havo- en vwo-examens meer (kortere) teksten en minder tekstbeschouwelijke vragen (in het bijzonder de functievragen). 'Niet-doorlopende teksten' komen vooralsnog minder frequent voor in de centrale examens van de hogere onderwijsniveaus dan bij PISA.

De centrale examens vmbo (gt en kb 2018) bevatten regelmatig vragen met een zuiver 'tekstbeschouwelijk' of 'tekstanalytisch' karakter (zonder de voor PISA kenmerkende reflecterende component). Om dergelijke vragen goed te kunnen beantwoorden, is specifiek onderricht en specifieke training noodzakelijk, zoals bijvoorbeeld over tekststructuur, opbouw van teksten en retorische middelen. Dergelijk onderwijs wordt vaak pas in klas drie en vaker nog in klas vier verzorgd. Vragen zoals in figuur 4.2 zouden in PISA-opgaven minder passen. De leesvaardigheid die bij PISA wordt gemeten, sluit meer aan bij het natuurlijke (alledaagse) lezen (waartoe ook het fictionele lezen behoort) en bij het alledaagse informatie verzamelen, verwerken en verstrekken. 'Leesvaardigheid' in de context van de centrale examens moet meer worden opgevat als (abstracte) tekstbeschouwing waarbij de analyse en interpretatie van gedetailleerde tekstkenmerken, van tekststructuren en van (complexe) argumentatie een rol spelen.

Tekst 1 De keerzijde van het spotprijsparadijs

1p **1** Een tekst kan op verschillende manieren ingeleid worden.
bijvoorbeeld door
1 een belangrijke conclusie voorop te stellen
2 een samenvatting van de tekst te geven
3 een voor de tekst belangrijke vraag te stellen
4 een voorbeeld bij het onderwerp te geven
Welke twee manieren worden in alinea's 1, 2 en 3 gebruikt om de tekst in te leiden?
A 1 en 2
B 1 en 4
C 2 en 3
D 2 en 4

Figuur 4.2 Vraag uit het examen vmbo-gl en tl 2018 (1e tijdvak) met een tekstanalytisch karakter

Een voorbeeld van een vraag uit het gt-examen 2018 die wél goed in een PISA-onderzoek zou passen, staat in figuur 4.3. Daarin wordt gevraagd naar de inhoud van een tekst, wat bij Pisa onder het proces Access and Retrieve valt. Dergelijke vragen hebben rechtstreeks betrekking op de inhoud van de tekst, zonder dat kennis nodig is over tekststructuren of dat gevraagd wordt naar een eigen mening of persoonlijke ervaringen in relatie tot deze tekst.

1p **3** In alinea 4 zeggen Lisa en Renée dat ze het eigenlijk afschuwelijk vinden om bij Primark te winkelen.
→ Noem de drie redenen die zij in alinea 4 geven waarom ze liever niet bij Primark winkelen.

Figuur 4.3 Vraag uit het Examen vmbo gl en tl 2018 (1e tijdvak) Nederlands CSE die ook in het PISA-onderzoek zou passen

Naar schatting 30% van de huidige eindexamenvragen is minder of niet vergelijkbaar met materiaal uit de PISA-afnames. Dergelijke vragen bevragen de structuur en tekstuele functies van tekstdelen, en het argumentatieve domein. Vragen en aangeboden teksten in de centrale examens zijn abstracter en complexer dan in PISA, en vergen vaak meer denkstappen. Maar dit is geen fundamenteel, maar slechts een gradueel verschil.

4.2 Wiskundige geletterdheid

4.2.1 Wat is wiskundige geletterdheid?

Wiskunde in het Nederlandse examenprogramma
Voor wiskunde is in het Nederlandse examen geen sprake van één vakdomein. Alleen al op havo/vwo was tussen 2006 en 2018 sprake van een grote hoeveelheid verschillende vakdomeinen: wiskunde A en B op havo en wiskunde A, B en C op vwo. Er is op havo/vwo ook nog het vak wiskunde D maar dat wordt alleen binnen het schoolexamen getoetst. Dat nog los van het feit dat de vakken wiskunde A, B en C in de loop van deze periode ook intern veranderden. Zo verviel in 2007 de heelvak-deelvakstruktuur in de centrale examens van wiskunde B havo/vwo en wiskunde A vwo.

In de periode 2006-2018 vonden verschillende veranderingen plaats in de wiskundeprogramma's havo/vwo. Ze deden hun intrede in het vierde jaar van havo/vwo, respectievelijk in 2007 (PEP) en 2015 (cTWO). De veranderingen werden doorgevoerd in de examens van:
- Examenjaar 2009: 1e afname havo-wiskunde A en B, PEP[11] (laatste afname havo-wiskunde A1,2, B1 en B1,2)[12]
- Examenjaar 2010: 1e afname vwo-wiskunde A, B en C, PEP (laatste afname vwo-wiskunde A1, A1,2, B1 en B1,2)
- Examenjaar 2017: 1e afname havo-wiskunde A en B, cTWO[13] (laatste afname havo-wiskunde A en B, PEP)
- Examenjaar 2018: 1e afname vwo-wiskunde A, B en C, cTWO (laatste afname vwo-wiskunde A, B en C, PEP)

11 PEP: Platform Examen Programma's

12 Van belang is nog op te merken dat er, voorafgaand aan de veranderingen die leidden tot de (her-)invoering van wiskunde A en B bij het havo-examenjaar 2009, in het havo-onderwijs ook het vak wiskunde A1 bestond. Dat vak werd niet op centraal-examenniveau getoetst maar enkel op schoolexamenniveau. Havo-wiskunde A1 kende wel een examenprogramma dat echter niet tot in dezelfde details gespecificeerd was als de andere, wel C.E.-getoetste wiskundeprogramma's.

13 cTWO: commissie Toekomst WiskundeOnderwijs

In de examenprogramma's vmbo is sinds 2006 niets veranderd. Wel waren er ontwikkelingen in de examinering. Zo is er sinds 2005 ook een digitaal examen voor vmbo bb, en sinds 2010 ook een digitaal examen voor vmbo kb. Tot en met 2006 werden statistiek en meetkunde afwisselend getoetst in vmbo-examens. Zo werd meetkunde in 2005 getoetst in het centraal examen vmbo, en statistiek alleen in het schoolexamen. In 2006 werd statistiek getoetst in het centraal examen, meetkunde alleen in het schoolexamen. In 2007 en later wordt meetkunde getoetst in het centraal examen, en statistiek in het schoolexamen. In 2017 is er voor het eerst geen overlap meer tussen de centrale examens van vmbo kb en gl/tl. Tot die tijd gold een overlap van ongeveer 40%.

Wiskunde in PISA

PISA kwam in 2012 met een herziene toets voor wiskunde. Het wiskunderaamwerk bevat vanaf dat moment een nieuwe formele definitie van wiskundige geletterdheid (zie tabel 4-4). Daarin worden vier inhoudelijke subdomeinen onderscheiden: Vorm en Ruimte, Veranderingen en Relaties, Onzekerheid en Hoeveelheid. Die wiskundige subdomeinen worden in PISA in verschillende contexten aangeboden: via vraagstukken in de persoonlijke levenssfeer, en in beroepsmatige, maatschappelijk gerelateerde en wetenschappelijk georiënteerde contexten.

> ***Definitie wiskundige geletterdheid in 2000/2003/2006/2009:***
> *Het vermogen van een individu om de rol die wiskunde speelt in de wereld, te kunnen identificeren en te begrijpen, het vermogen om gefundeerde beslissingen te nemen en om wiskunde te gebruiken op een wijze die tegemoet komt aan de behoeften van diens leven als een opbouwend, betrokken en beschouwend burger.*

> ***Definitie wiskundige geletterdheid in 2012/2015:***
> *Wiskundige geletterdheid is het vermogen van een individu om wiskunde in een diversiteit van contexten te formuleren, gebruiken en interpreteren. Het bevat wiskundig redeneren en het gebruiken van wiskundige concepten, procedures, kennis en instrumenten waarmee verschijnselen beschreven, verklaard en voorspeld kunnen worden. Het helpt individuen de rol die wiskunde speelt in de wereld te herkennen en goed doordachte oordelen en beslissingen te nemen die noodzakelijk zijn voor opbouwende, betrokken en beschouwende burgers.*

In 2012 werden ook drie nieuwe schalen ontwikkeld voor de meting van de wiskundige competenties. De gemeten competenties waarmee een wiskundevraagstuk aangepakt dient te worden, zijn 'formuleren', 'toepassen' en 'interpreteren'. Om de scores van verschillende jaren op dezelfde schaal te kunnen weergeven, bevat de toets uit 2012 een aantal gemeenschappelijke opgaven (de ankeropgaven) met de wiskundetoetsen uit 2003 (en 2006, 2009 en ook 2000).

Vergelijken we de oude en herziene definitie, dan kunnen de opgaven uit de PISA-toets van 2003 ook nog gebruikt worden in de PISA-toetsen tot en met 2015. In het geval van de ankeropgaven is dat ook daadwerkelijk gebeurd. De herziene definitie heeft dan ook meer te maken met het anders formuleren van bepaalde aspecten dan met een veranderde visie op wiskunde/wiskundige geletterdheid. Bij de herziening werd wel een dimensie toegevoegd, maar ook die zorgt er niet voor dat bepaalde oude PISA-vragen niet meer gesteld mogen worden of bepaalde nieuwe PISA-vragen in het verleden niet gesteld hadden kunnen worden.

Tabel 4.4 Kenmerken en taakaspecten wiskundige geletterdheid PISA 2012 (Cito, 2013, p. 16)

Definitie en belangrijkste kenmerken	• Wiskundige geletterdheid is het vermogen van een individu om wiskunde in een diversiteit van contexten te formuleren, gebruiken en interpreteren. • Het bevat wiskundig redeneren en het gebruiken van wiskundige concepten, procedures, kennis en instrumenten waarmee verschijnselen beschreven, verklaard en voorspeld kunnen worden. • Het helpt individuen de rol die wiskunde speelt in de wereld te herkennen en goed doordachte oordelen en beslissingen te nemen die noodzakelijk zijn voor opbouwende, betrokken en beschouwende burgers.
Kennisdomein	• Vorm en Ruimte • Veranderingen en Relaties • Onzekerheid • Hoeveelheid
Relevante competenties	• Formuleren • Toepassen • Interpreteren
Context en situatie	De toepassingsgebieden van de wiskunde: • Persoonlijk • Schools en beroepsmatig • Publiek • Wetenschappelijk

Overlap en verschillen

Vrijwel alles wat er in de syllabi van alle wiskunde-examenprogramma's havo/vwo stond en staat, kan nadrukkelijk niet onder de PISA-definities vallen. Deze examenprogramma's zijn gericht op onderwijs in de laatste paar jaren van het voortgezet onderwijs, en bevatten diepgaande wiskundige kennis en vaardigheden waaraan een leerling van 15 jaar redelijkerwijs niet 'blootgesteld' kan en mag worden.

Andersom is alles in de PISA-definitie onderliggend (c.q. voorbereidend) op hetgeen vermeld staat in de wiskunde-examenprogramma's havo/vwo. De PISA-toets is curriculumonafhankelijk. Het curriculum is in het examenprogramma nauwkeurig gespecificeerd en daar wordt in de centrale examens specifiek op getoetst. Bij PISA worden deze elementen juist niet getoetst. PISA neemt juist de algemene dagelijkse praktijk als basis. Dat wat bij PISA vermeld wordt, is ook niet makkelijk te herkennen in de wiskunde-examenprogramma's havo/vwo. Die examenprogramma's melden bijvoorbeeld niet dat verondersteld wordt dat leerlingen de rekentafels beheersen. Dat is zo basaal, dat daarop in examenprogramma's zonder vermelding op wordt voortgeborduurd. Om toch een indruk te krijgen van potentiële overlap hebben we de specificaties van vaardigheden in twee syllabi doorgenomen. Die specificaties blijken grotendeel over technische vaardigheden en over kennis te gaan die niet in PISA getoetst worden. Voorbeelden zijn het algebraïsch oplossen van vergelijkingen, het construeren van raaklijnen aan een grafiek, kennis van exponentiële en periodieke functies, en rekenen met coördinaten in een assenstelsel.

4.2.2 Hoe wordt wiskundige geletterdheid getoetst?

Bij PISA is het overgrote deel van de opgaven van een context voorzien. Daarmee lijkt PISA veel meer in het voortraject van de contextrijke vakken wiskunde A en C te passen. Maar ook bij wiskunde B bevatten het onderwijs en de opgaven voor een deel een context. In het vmbo vormen contexten meestal het uitgangspunt voor de examens. Tot slot stelt bevat PISA ook mondjesmaat intramathematische opgaven, opgaven waarbij de buitenwiskundige werkelijkheid geen rol speelt.

Bij PISA is een substantieel deel van de vragen van het gesloten type. Ook kort-antwoordvragen zijn fors vertegenwoordigd. Beide vraagtypen komen bij examens wiskunde niet of uiterst zelden voor. Ook de scoring bij PISA is van een andere orde dan die van examens. Bij de beoordeling van PISA-vragen wordt, grofweg, meer gelet op de geest van de gedachtegang dan op de concrete uitvoering. Als helder is dat een leerling het bevraagde concept doorgrond heeft, maar het (weinige) rekenwerk niet (helemaal) goed heeft uitgevoerd, kan deze toch de maximale code (want zo heet dat bij PISA) ontvangen. Bij de Nederlandse examencorrectie is een dergelijke benadering niet bepaald standaard. Daar komt bij dat bij PISA twee min of meer vergelijkbare vragen soms essentieel verschillend gecodeerd moeten worden: bij de ene vraag kan dan veel sneller een maximale code gegeven worden dan bij de andere. Die laatste vraag is dan moeilijker, omdat het hoogst haalbare eindstation ‘verder’ weg ligt.

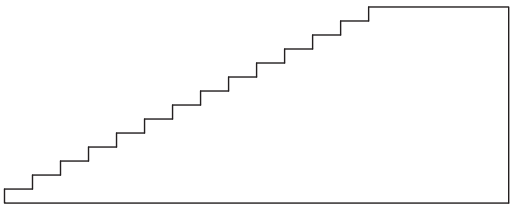
Als we ons beperken tot havo/vwo is er geen enkele PISA-opgave die zou passen in een examen havo/vwo (voor welke wiskunde dan ook). Daarbij baseren we ons in eerste instantie toch op niveau, want op grond daarvan worden onze examens nu eenmaal gemaakt. PISA-vragen bevatten zo weinig van in de laatste jaren onderwezen wiskunde, dat het stellen van een PISA-vraag in een examen onmiddellijk tot ophef in het veld zou leiden. Omgekeerd geldt min of meer hetzelfde: in havo/vwo-examens is geen enkele opgave te vinden die binnen PISA gesteld zou kunnen worden.

Tot slot verwijzen we naar het rapport van PISA 2003. Dit geeft diverse voorbeeldopgaven voor wiskundige geletterdheid (Bijlage 4 Voorbeeldopgaven wiskunde, pp. 125-164) en voorziet ze van een moeilijkheidsscore op de PISA-schaal (tabel 2.6, p. 41). Ook wordt hier in generieke, maar ook domeinspecifieke termen vermeld welke vaardigheden bij welke vaardigheidsscore horen (tabel 2.3, pp. 29-30). De voorbeelden illustreren welke wiskundige vaardigheden PISA aan de orde stelt.

TRAP

Vraag 1: TRAP M547Q01

Hieronder zie je de afbeelding van een trap met 14 treden met een totale hoogte van 252 cm:



Wat is de hoogte van elk van de 14 treden?

Hoogte: .....cm.

Figuur 4.4 Voorbeeldopgave uit PISA 2003 uit het kennisdomein Vorm en Ruimte met een moeilijkheidsniveau 2 (Cito, 2004)

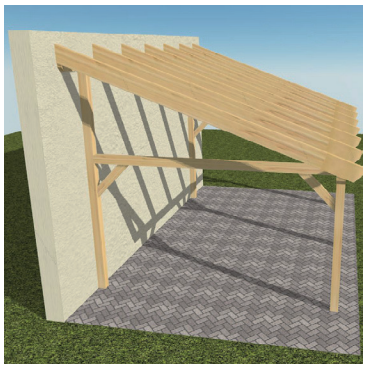
Figuur 4.4 bevat een voorbeeldopgave. Deze opgave valt volgens de beschrijving onder het kennisdomein Vorm en Ruimte en heeft een moeilijkheidsniveau 2. Op niveau 2 kan een leerling (Cito, 2004, pp. 29-30):

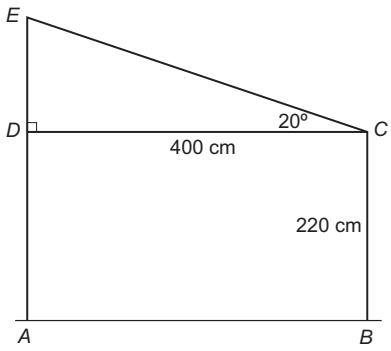
- situaties in contexten interpreteren en herkennen op basis van directe gevolgtrekkingen;
- relevante informatie onttrekken aan een enkele bron;
- gebruikmaken van een enkele representatievorm;
- gebruikmaken van elementaire algoritmes, formules, procedures of afspraken;
- gebruikmaken van eenvoudig redeneren;
- letterlijke interpretaties maken van resultaten.

Ter vergelijking bevat figuur 4.5 een opgave uit het centraal schriftelijk eindexamen wiskunde vmbo-gt in 2018. Ook in deze examenopgave wordt gevraagd om een hoogte. Op het eerste gezicht lijkt dat op vraag 1 van de PISA-opgave TRAP. Maar om de examenvraag te kunnen beantwoorden, moet de leerling beschikken over kennis van goniometrie. Die kennis is onderdeel van het vmbo-gl/tl-programma, en hoort niet tot het wiskundige instrumentarium van leerlingen die de PISA-opgaven maken.

Carport

Carin wil een carport bouwen. Voordat ze begint, maakt ze eerst op de computer een ontwerp en een schets van het vooraanzicht.





vooraanzicht

In de schets van het vooraanzicht is de breedte 4 meter en de inrijhoogte 2,2 meter. Hoek *C* = 20°.

4p **11** Bereken, zonder te meten, hoeveel cm de hoogte *AE* is. Schrijf je berekening op.

Figuur 4.5 Vraag uit het Examen vmbo gl en tl 2018 (2e tijdvak) Wiskunde CSE

4.3 Natuurwetenschappelijke geletterdheid

4.3.1 Wat is natuurwetenschappelijke geletterdheid?

Natuurwetenschappen in de Nederlandse examenprogramma's

In het programma van de centrale examens komt het vak natuurwetenschappen als zodanig niet voor. In Nederland hebben we het over drie verschillende vakken (aardrijkskunde, biologie en natuurkunde), elk op drie niveaus (vmbo, havo, vwo). Voor de volledigheid zou hier ook nask-2 en scheikunde genoemd moeten worden. De drie domeinen zijn globaal beschreven in de examenprogramma's en nader uitgewerkt in de syllabi. Elke syllabus geeft een beschrijving van de algemene vaardigheden, de vakvaardigheden en de inhoudelijke domeinen. Zeker bij natuurkunde en biologie zijn die contextgebonden beschreven, met daarbij de toepassingsgebieden die bekend worden verondersteld.

Natuurwetenschappen in PISA

Om te benadrukken dat in PISA het toepassen van natuurwetenschappelijke kennis in alledaagse situaties belangrijk is, gebruikt PISA de term 'natuurwetenschappelijke geletterdheid'. Dit wordt opgehangen aan de volgende competenties:

- fenomenen natuurwetenschappelijk *verklaren*;
- natuurwetenschappelijk onderzoek *evalueren en ontwerpen*;
- gegevens en bewijzen natuurwetenschappelijk *interpreteren*.

Belangrijk is dat het gaat om algemene natuurwetenschappelijke kennis en vaardigheden. PISA is niet of nauwelijks curriculumgebonden. Dit maakt het mogelijk resultaten van de verschillende landen met elkaar te vergelijken. In 2015 werd de definitie van natuurwetenschappelijke geletterdheid grondig herzien.

Definitie natuurwetenschappelijke geletterdheid in 2006/2009/2012:
Natuurwetenschappelijke kennis en gebruik van die kennis om problemen te herkennen, nieuwe kennis op te doen, natuurwetenschappelijke verschijnselen te verklaren, en gefundeerde conclusies te trekken betreffende onderwerpen met een natuurwetenschappelijke inhoud.

- *Inzicht in karakteristieke kenmerken van de natuurwetenschappen en hoe deze zijn te herkennen in onderzoek en kennisontwikkeling.*
- *Begrip van de rol die natuurwetenschappen, techniek en technologie spelen bij de vorming van onze materiële, intellectuele en culturele omgeving.*
- *Bereidheid om zich als weldenkend burger te verdiepen in onderwerpen en opvattingen met een natuurwetenschappelijke inhoud.*

Definitie natuurwetenschappelijke geletterdheid in 2015:
Het vermogen om zich bezig te houden met natuurwetenschappelijke kwesties en met natuurwetenschappelijke ideeën als weldenkend burger. Een natuurwetenschappelijk geletterde persoon is bereid om een beredeneerd betoog over natuurwetenschap en technologie te houden, waarvoor de volgende competenties nodig zijn:

- *De competentie om fenomenen natuurwetenschappelijk te verklaren (het herkennen, geven en evalueren van verklaringen voor een reeks natuurlijke en technologische verschijnselen).*
- *De competentie om natuurwetenschappelijk onderzoek te evalueren en te ontwerpen (het beschrijven en evalueren van natuurwetenschappelijk onderzoek en manieren voorstellen om vraagstukken natuurwetenschappelijk te beantwoorden).*
- *De competentie om gegevens en bewijs natuurwetenschappelijk te interpreteren (het analyseren en evalueren van gegevens, beweringen en argumenten in verschillende vormen en het trekken van gepaste natuurwetenschappelijke conclusies).*

De vragen zijn (ongeveer gelijk) verdeeld over de drie vakgebieden ‘levende natuur’, ‘niet-levende natuur’ en ‘aarde en ruimte’. In deze verdeling herkennen we onderdelen uit de schoolvakken Biologie, Natuurkunde/Scheikunde en Fysische Geografie (zie tabel 4.5.)

De vakkennis bij natuurwetenschappelijke geletterdheid voldoet aan de volgende criteria:

- relevantie voor het dagelijks leven;
- representativiteit voor belangrijke natuurwetenschappelijke concepten en als zodanig langdurig bruikbaar;
- geschiktheid voor het kennisniveau van 15-jarigen.

Volgend uit de definitie van natuurwetenschappelijke geletterdheid, stelt PISA de vragen altijd in een van de volgende contexten: gezondheid, natuurlijke hulpbronnen, milieu, risico’s en nieuwe ontwikkelingen in natuurwetenschap. De vragen hebben betrekking op persoonlijke, nationale en globale kwesties, zowel hedendaags als historisch, die begrip van natuurwetenschap en technologie vereisen.

De subdomeinen die PISA bevraagt, werden in 2015, samen met de definitie, gewijzigd. In 2006 werden twee subdomeinen onderscheiden: Kennis van natuurwetenschappen (Vakkennis) en Kennis over natuurwetenschappen. In 2015 heten de twee subdomeinen Vakkennis en Kennisvorming. Bij kennisvorming gaat het om kennis over de methoden van natuurwetenschappelijk onderzoek (procedurele kennis) en kennis over de rationale achter deze methoden en de rechtvaardiging van het gebruik ervan (epistemische kennis). Het gevolg is dat er geen een-op-een relatie meer is tussen de subdomeinen van natuurwetenschappen in 2015 en de subdomeinen bij eerdere afnames.

Tabel 4.5 Kenmerken en taakaspecten natuurwetenschappelijke geletterdheid PISA 2015

Definitie en belangrijkste kenmerken	• Het vermogen om zich bezig te houden met natuurwetenschappelijke kwesties en met natuurwetenschappelijke ideeën als weldenkend burger. • Inzicht in karakteristieke kenmerken van de natuurwetenschappen en hoe deze zijn te herkennen in onderzoek en kennisontwikkeling. • Begrip van de rol die natuurwetenschappen, techniek en technologie spelen bij de vorming van onze materiële, intellectuele en culturele omgeving. • Bereidheid om zich als weldenkend burger te verdiepen in onderwerpen en opvattingen met een natuurwetenschappelijke inhoud.
Kennisdomein	• Levende natuur • Aarde en ruimte • Niet-levende natuur
Relevante competenties	• Verschijnselen natuurwetenschappelijk verklaren • Evalueren en ontwerpen van natuurwetenschappelijk onderzoek • Natuurwetenschappelijk interpreteren van gegevens en bewijzen
Context en situatie	De toepassingsgebieden van de natuurwetenschappen: • Gezondheid • Natuurlijke hulpbronnen • Milieu • Risico's • Grenzen van natuurwetenschappen en techniek

Overlap en verschillen

Vrijwel niets uit de syllabi van de natuurwetenschappelijke vakken havo/vwo kan onder de PISA-definities vallen. Onze examenprogramma's zijn gericht op onderwijs in de laatste paar jaren van het voortgezet onderwijs en bevatten dan ook kennis en vaardigheden van een zodanige diepgang dat die niet gevraagd kan worden van een leerling van 15 jaar. Hooguit een vergelijking met vmbo-4 heeft zin. PISA richt zich op algemene inzichten en toepassingen van kennis en vaardigheden aan het eind van de onderbouw. In de onderbouw wordt de basis gelegd voor de kennis die leerlingen die het betreffende vak kiezen, uitdiepen en uitbreiden in de bovenbouw. Daarnaast is de inhoud van wat in de onderbouw aan de orde komt slechts globaal vastgelegd. Scholen hebben grote vrijheid in de keuze voor het aantal lessen en de vakinhoud. Dat maakt een vergelijking tussen PISA en het domein ingewikkeld.

Waar PISA en de centrale examens wel overeenkomen, is dat de opgaven contextgebonden zijn. De vragen zijn geclusterd in opgaven die een situatie uit het dagelijks leven (of de natuurwetenschap) behandelen. De keuze van de contexten verschilt echter. In PISA staat alles in het teken van burgerschap. De achtergrond daarbij is dat je als burger natuurwetenschappelijke geletterdheid nodig hebt om te volgen wat er speelt en tot keuzes te komen op persoonlijk, landelijk en globaal niveau (bijvoorbeeld voor de aanschaf van apparaten, het doen van investeringen, de landelijke verkiezingen). In de eindexamens staat de toets in het teken van vakmanschap. Deze kennis gaat meestal verder dan de kennis voor goed burgerschap. Zo gaat het in de eindexamens niet alleen om kennis nodig voor het lezen van de krant of een folder, maar om een apparaat te ontwerpen of repareren. Het gaat ook niet alleen om het kunnen kiezen van een goede verpleger of dokter, maar om een goede verpleger of dokter te zijn. De eindexamencontexten bestrijken dus een veel groter gebied dan de PISA-contexten.

4.3.2 Hoe wordt natuurwetenschappelijke geletterdheid getoetst?

In 2015 is PISA overgestapt van papieren toetsen naar toetsen op de computer. Daarvoor zijn oude papieren opgaven omgezet naar computeropgaven, maar er werden ook nieuwe opgaven geconstrueerd (onder andere door Cito-medewerkers). In deze nieuwe opgaven wordt gebruik gemaakt van de extra mogelijkheden die de computer biedt: kleuren, filmpjes en interactieve animaties, waarbij leerlingen keuzes moeten maken en handelingen moeten verrichten. Bij de centrale examens wordt alleen in het vmbo (basis en kader) getoetst met de computer. Dergelijke interactieve applicaties zijn daar echter zeldzaam. In vmbo gl/tl-, havo- en vwo-examens wordt nog geen gebruik gemaakt van examens op de computer.

Vragen zijn in PISA gesloten of van het type 'kort antwoord'. Dat geldt ook voor de interactieve animaties. In onze centrale examens vmbo is een gedeelte van de vragen gesloten en voorgestructureerd zoals bij PISA. In havo en vwo worden bij biologie ook gesloten vragen gebruikt, maar bij natuurkunde worden alleen bij havo incidenteel gesloten vragen gebruikt en bij aardrijkskunde en scheikunde helemaal niet.

De scoring bij PISA is van een andere orde dan die van de vmbo-examens. Grofweg let PISA meer op de geest van de gedachtegang dan op de concrete formulering. Als duidelijk is dat een leerling het bevraagde concept heeft doorgrond, maar het rekenwerk niet (helemaal) goed heeft uitgevoerd, wordt soms toch de maximale score toegekend. Bij de centrale examens is de score voor een niet-gesloten vraag opgedeeld in deelscores, die onafhankelijk van elkaar te behalen zijn. Terwijl het aantal deelscorepunten bij examens iets zegt over de complexiteit van de vraag, is dat bij PISA niet het geval.

Waar PISA en de centrale examens overeenkomen, is dat er nauwelijks naar feitenkennis wordt gevraagd. Informatie wordt gegeven in de opgaven; leerlingen moeten deze interpreteren. Verschil is wel dat PISA-opgaven vaak zeer veel tekst bevatten. Inhoudelijk zouden complete PISA-opdrachten niet in een eindexamen vmbo passen. Daarvoor zijn ze te veel gericht op algemene kennis, zoals verwoord in het begrip 'geletterdheid'. Figuur 4.6 bevat een voorbeeld: de opdracht gaat in op procedurele kennis over onderzoek doen, en dat valt niet onder het vmbo-examenprogramma. Sommige losse vragen uit een PISA-opgave zouden wel in een vmbo-examen kunnen zitten. Dat geldt bijvoorbeeld voor de voorbeeldvraag in figuur 4.7.

NIVEAU 2

Vraag 3: GENETISCH GEMODIFICEERDE GEWASSEN S508Q03

Er is maïs geplant op 200 akkers verspreid over het land. Waarom hebben de wetenschappers dat op meer dan één plaats gedaan?

A Zodat veel landbouwers het nieuwe GM-maïs konden proberen.
B Om te kijken hoeveel GM-maïs ze konden verbouwen.
C Om zo veel mogelijk land te bedekken met het GM-gewas.
D Om er verschillende groeiomstandigheden voor maïs bij te betrekken.

Figuur 4.6 Voorbeeld van een losse vraag uit een voorbeeldopgave van PISA 2006 die niet in een examen vmbo zou passen

LICHAAMSBEWEGING

Regelmatige lichaamsbeweging is goed voor de gezondheid, als het maar met mate gebeurt.

Vraag 5: LICHAAMSBEWEGING S493Q05 – 01 11 12 99

Waarom moet je sneller en dieper ademhalen als je aan lichaamsbeweging doet, dan wanneer je lichaam in rust is?

Figuur 4.7 Voorbeeld van een losse vraag uit een voorbeeldopgave van PISA 2006 die ook in een examen vmbo zou passen

Of eindexamenopgaven in PISA zouden passen, wordt duidelijk uit een vergelijking van de examens vmbo gl/tl uit 2018. Soms bevatten deze examens een vraag die ook in PISA gesteld zou kunnen worden, meestal als onderdeel van een opgave die ook vragen over de vakinhoud bevat. Het gaat om de volgende aantallen vragen:

- Vier van de 45 vragen (9%) in het examen aardrijkskunde vmbo gl/tl 2018-1. Eén vraag gaat over het ontstaan van het broeikaseffect (zie figuur 4.8). Bij de andere drie vragen zijn bronnen gegeven, waar de leerlingen conclusies uit moeten trekken.
- Eén van de 46 vragen (2%) in het examen aardrijkskunde vmbo gl/tl 2018-2. Bij deze vraag moeten conclusies getrokken worden uit een krantenartikel.
- Twee van de 54 vragen (3%) in het examen biologie vmbo gl/tl 2018-1. Bij deze vragen wordt een onderzoek beschreven, waar leerlingen conclusies uit moeten trekken.
- Eén van de 49 vragen (2%) in het examen biologie vmbo gl/tl 2018-2. Bij deze vraag moeten conclusies worden getrokken uit de omschrijving van een onderzoek.
- Geen enkele vraag (0%) in het examen nask-1 vmbo gl/tl 2018-1.
- Geen enkele vraag (0%) in het examen nask-1 vmbo gl/tl 2018-2.

Weer en klimaat

1p **1** In de volgende drie stappen wordt het ontstaan van het broeikaseffect beschreven. De stappen staan in willekeurige volgorde.

1 De aarde geeft warmte van de weerkaatste zonnestralen af aan de atmosfeer.
2 De binnenkomende zonnestralen gaan door de atmosfeer heen naar de aarde.
3 De gemiddelde temperatuur op aarde stijgt, omdat een deel van de zonnestralen teruggekaatst wordt naar de aarde.

Wat is de juiste volgorde van het ontstaan van het broeikaseffect?

A 1 - 2 - 3
B 1 - 3 - 2
C 2 - 1 - 3
D 2 - 3 - 1

Figuur 4.8 Vraag uit het Examen vmbo gl en tl 2018 (1e tijdvak) Aardrijkskunde CSE die ook in PISA zou passen

4.4 Samenvatting en conclusie

In de hoofdstuk hebben we onderzocht welke verschillen en overeenkomsten er zijn tussen de inhoud van PISA-domeinen en het examenprogramma. Welke verschillen en overeenkomsten zijn er tussen de domeinbeschrijving en opgaven van PISA's Leesvaardigheid, Wiskundige geletterdheid en Natuurwetenschappelijke geletterdheid en de inhoud van het examenprogramma in Nederland?

Voor leesvaardigheid zien we duidelijke overeenkomsten tussen de PISA-definitie en de omschrijving van leesvaardigheid in de eindtermen vmbo. Tegelijkertijd is het evident dat het vakdomein Nederlandse taal veel meer omvat dan leesvaardigheid. Terwijl sommige PISA-vragen zouden kunnen voorkomen in de vmbo-examens, kan een groot deel van de examenvragen niet voorkomen in PISA, met name vanwege het tekstbeschouwelijke karakter. De examenvragen zijn ook abstracter en complexer.

Voor wiskundige geletterdheid in PISA en de verschillende wiskunde-vakdomeinen in de centrale examens geldt dat de overeenkomsten gering zijn, met name op havo- en vwo-niveau. Wiskundige geletterdheid is voorbereidend op het examenprogramma, maar zo basaal dat het moeilijk te herkennen is in de diverse examenprogramma's. Ook de vragen in beide toetsen zijn verschillend. Zo zijn in PISA de meeste vragen gesloten, wat in het examen niet het geval is. Ook is de scoring in PISA heel coulant.

In de natuurwetenschappelijke geletterdheid van PISA zijn de vakken aardrijkskunde, biologie en natuurkunde te herkennen, maar is er echt een verschil tussen geletterdheid en vakkennis. Qua vragen zien we de (kleine) overeenkomst dat ze contextgebonden zijn. De keuze van deze contexten is echter geheel anders: waar PISA-contexten in het teken staan van burgerschap, staan de contexten in de examens in het teken van vakmanschap. De verschillen tussen natuurwetenschappelijke geletterdheid en de vakken aardrijkskunde, biologie en natuurkunde zijn zo groot dat een zinvolle vergelijking niet mogelijk is.

5 Leerlingenstromen

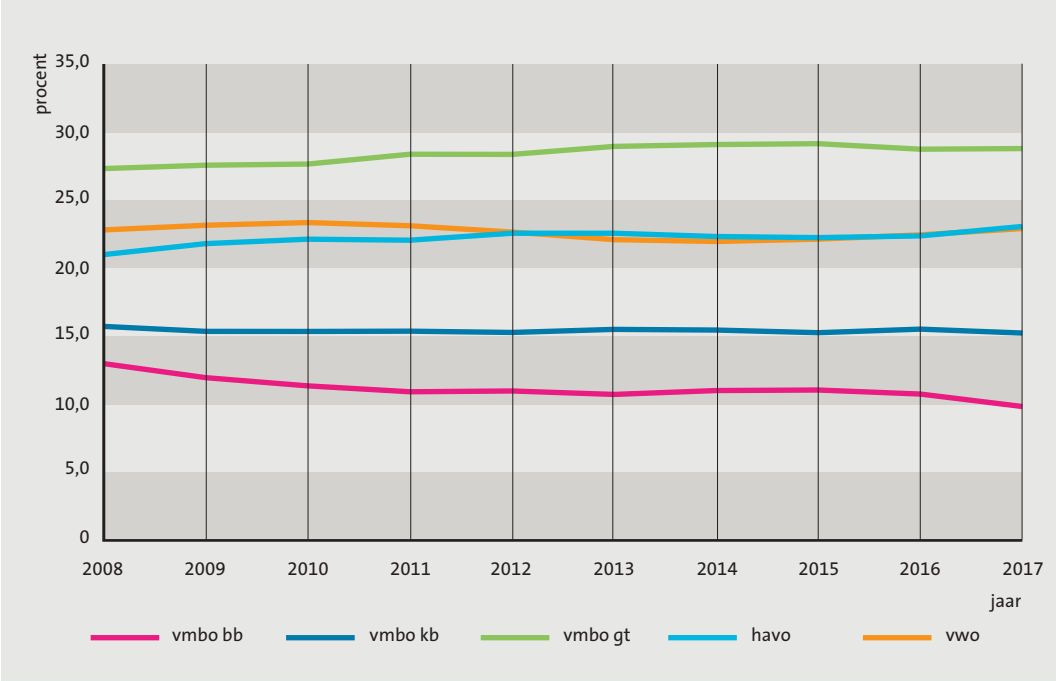
5 Leerlingenstromen

Eén verklarende factor voor de geconstateerde verschillen tussen PISA en het eindexamen kan een verschil in de leerlingpopulaties zijn. PISA wordt internationaal afgenomen onder een – voor elk land – representatieve groep 15-jarigen. In Nederland zitten deze leerlingen aan het einde van de onderbouw of begin van de bovenbouw. Qua leeftijd en genoten onderwijs (en dus kennis en vaardigheden) komen deze grotendeels niet overeen met de leerlingen die in Nederland eindexamen doen. Globaal gezien lijkt de PISA-populatie nog het meest op de populatie examenleerlingen-vmbo. Dat maakt over en weer gemeten vaardigheden lastig vergelijkbaar.

In de jaren tussen einde onderbouw en het eindexamen (einde bovenbouw) genieten de leerlingen nog één tot drie jaar onderwijs. Maar ook andere factoren dan het genoten onderwijs kunnen in die periode ertoe leiden dat leerlingen met een hoger vaardigheidsniveau aan de eindexamens beginnen. Zoals uit hoofdstuk 2 al bleek, lijken de aangescherpte exameneisen in Nederland hun invloed te hebben gehad op de gemiddelde cijfers van het centraal examen.
In 2011 was sprake van een keerpunt: de dalende trend in de cijferontwikkeling sloeg in 2012 om in een stijgende trend. Het is mogelijk dat scholen in aanloop of in reactie op die strengere exameneisen maatregelen hebben genomen die van invloed zijn geweest op het vaardigheidsniveau. Ze kunnen bijvoorbeeld strengere overgangsnormen hanteren of leerlingen sneller laten afstromen.
DUO houdt in de Examenmonitor de verdeling van het aantal leerlingen over de verschillende onderwijsniveaus bij. Ook doorstroomfactoren in de bovenbouw die van invloed kunnen zijn op de examenresultaten (zoals zittenblijven en afstromen), worden geregistreerd. We bekijken ze in dit hoofdstuk.

5.1 Verdeling leerlingen over schooltypen en leerwegen

Figuur 5.1 laat over de laatste tien jaar de verdeling van bovenbouwleerlingen over de verschillende schooltypen en leerwegen zien. Duidelijk is dat die verdeling licht fluctueert. Over het geheel genomen neemt het percentage leerlingen in vmbo-bb af (-3,2%). Daar staat tegenover dat het percentage leerlingen in vmbo-gt en havo stijgt (respectievelijk +1,5% en +2,1%). Het percentage leerlingen in vmbo-kb blijft behoorlijk constant (-0,5%), en het percentage vwo-leerlingen is in 2017 weer vrijwel gelijk aan dat van 2008 (+0,1%).



Figuur 5.1 Verdeling leerlingen over schoolsoorten en leerwegen in het derde leerjaar (bovenbouw vmbo, havo en vwo)

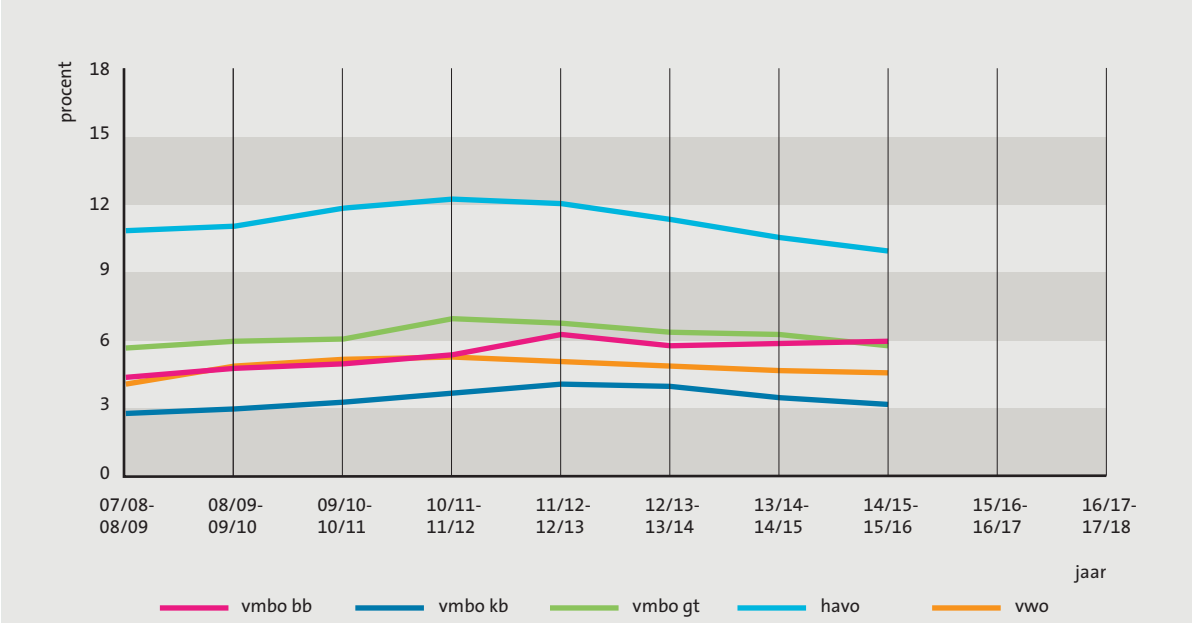
Bron: https://www.onderwijsincijfers.nl/kengetallen/voortgezet-onderwijs/deelnemersvo/leerlingen-in-het-derde-leerjaar

5.2 Doorstroom in de bovenbouw

Zittenblijven in de bovenbouw

Hoewel de Examenmonitor van 2017 geen gegevens over zittenblijven geeft, geeft de Examenmonitor van 2016 (DUO, 2016) wel informatie over het aantal zittenblijvers in de bovenbouw[14] (exclusief het examenjaar). Figuur 5.2 maakt duidelijk dat het aantal zittenblijvers onder havoleerlingen aanzienlijk hoger is dan op andere schoolsoorten. Elk jaar is het aantal zittenblijvers in havo ongeveer twee keer zo hoog als in andere schoolsoorten. Focussen we op de aanloop naar de strengere exameneisen in 2012, dan stijgt het aantal zittenblijvers over de hele linie licht. Daarna neemt het weer geleidelijk af, behalve in vmbo-bb. De piek in zittenblijvers ligt in vmbo-gt, havo en vwo tussen schooljaar 10/11 en 11/12, voor vmbo-bb en vmbo-kb een jaar later.

14 Vmbo: leerjaar 3; havo: leerjaar 3 en 4; vwo: leerjaar 3, 4 en 5.

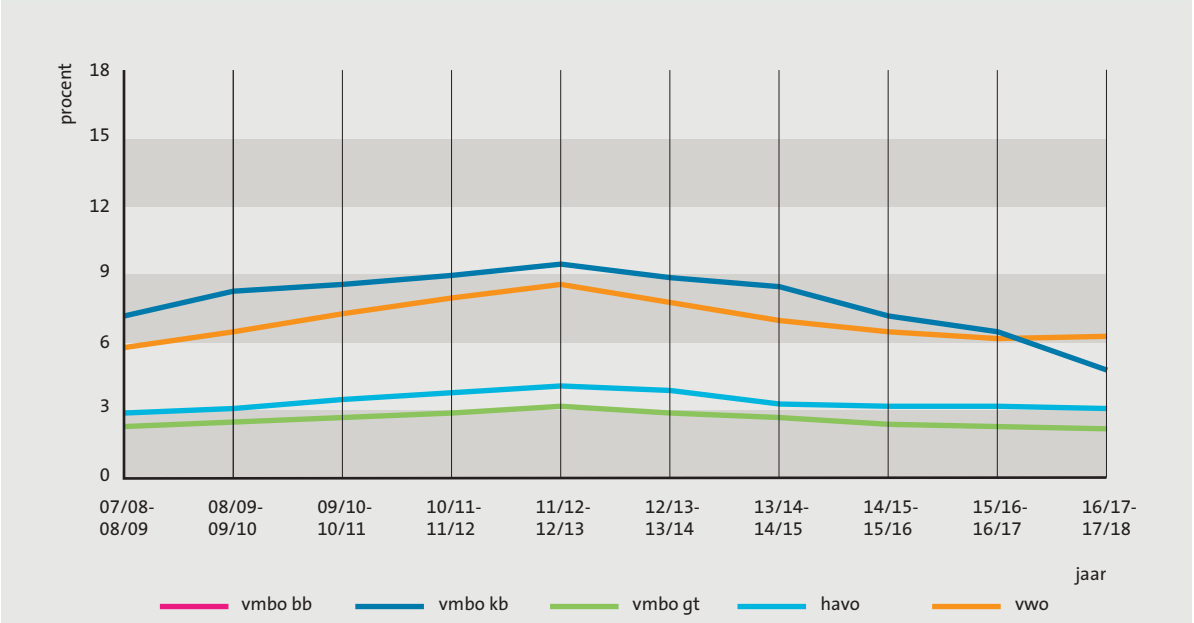


	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Rekentoets; Exameneis Nederlands aangescherpt	CE Neder-lands geijkt; (Rekentoets)	(Rekentoets)	(Rekentoets)
Havo/Vwo		Heelvak/ deelvak afgechaft (Havo)	Heelvak/ deelvak afgeschaft (Vwo)		Aange-scherpte exameneis	Kernvakken-regeling	Rekentoets	CE Neder-lands geijkt & geleide samenvatting afgeschaft; (Rekentoets) Nieuw examen-programma Natuurweten-schappen (Havo)	Rekentoets telt mee in examenresul-taat (Vwo) Nieuw examen-programma Natuurweten-schappen (Vwo)	Rekentoets telt mee in kernvakken-regel (Vwo) Nieuw examen-programma Wiskunde (Havo)

Figuur 5.2 Percentage zittenblijvers in de bovenbouw (exclusief het examenjaar) per schoolsoort

Afstroom

De vraag is of het beeld dat wordt neergezet door het zittenblijfpercentage ook terugkomt in het afstroompercentage. Het gaat dan om leerlingen die na een jaarovergang het onderwijs vervolgen in een lagere leerweg of een lager schooltype. Figuur 5.3 geeft over de laatste tien jaar inzicht in het percentage afgestroomde leerlingen. Net zoals bij het aantal zittenblijven neemt de afstroom toe in aanloop naar de scherpere exameneisen. Daarna daalt het afstroompercentage weer geleidelijk.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Rekentoets; Exameneis Nederlands aangescherpt	CE Neder-lands geijkt; (Reken-toets)	(Rekentoets)	(Rekentoets)
Havo/Vwo		Heelvak/deel-vak af-geschaft (Havo)	Heelvak/deel-vak af-geschaft (Vwo)		Aange-scherpte exameneis	Kernvakken-regeling	Rekentoets	CE Neder-lands geijkt & geleide samenvatting afgeschaft; (Rekentoets)	Rekentoets telt mee in examenresul-taat (Vwo)	Rekentoets telt mee in kernvakken-regel (Vwo)
								Nieuw examen-programma Natuurweten-schappen (Havo)	Nieuw examenpro-gramma Natuurweten-schappen (Vwo)	Nieuw examenpro-gramma Wis-kunde (Havo)

Figuur 5.3 Percentage afgestroomde leerlingen in de bovenbouw (exclusief het examenjaar)

Zowel een toename van zittenblijven als afstroom zou pas één tot drie schooljaren later tot een stijging van de vaardigheid in het examenjaar (examencijfers) kunnen leiden. De afstroom is bepaald in de bovenbouw waarbij de examenklassen niet zijn meegerekend. In het examenjaar komt vrijwel geen afstroom voor. In het vmbo gaat het dus om de afstroom in leerjaar 3, in het havo om de gemiddelde afstroom in de leerjaren 3-4, en in het vwo om de leerjaren 3-5. Stel: bij de meting in oktober 2008 (schooljaar 2008/2009) blijkt dat een aantal leerlingen in leerjaar 3 is blijven zitten of is afgestroomd (vanuit schooljaar 2007/2008). Deze leerlingen doen dan op zijn vroegst het volgende schooljaar eindexamen (examenjaar 2010).
In tabel 5.1 is de correlatie berekend tussen het percentage zittenblijvers en afstromers aan het

begin van een schooljaar (x) en de examenresultaten in het volgende schooljaar (examenjaar x+2). Er blijkt dan een positieve samenhang te zijn. Een toename van zittenblijven en afstromen, hangt samen met hogere examencijfers een jaar later. (NB. Het effect van zittenblijven en afstromen zou exacter te bepalen zijn, als van examenkandidaten bekend was of het gaat om zittenblijvers of afstromers uit hogere leerwegen/schooltypen).

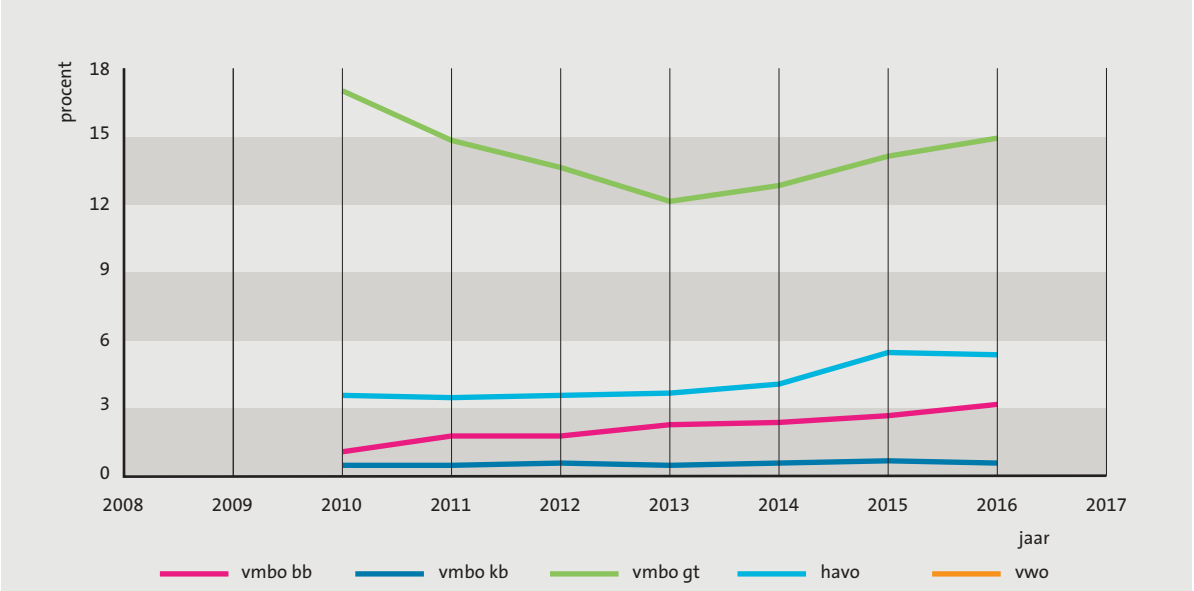
	Samenhang tussen afstroom uit hogere leerweg en cijfer CE	Samenhang tussen afstroom en cijfer CE lagere leerweg/schooltype	Samenhang tussen zittenblijven en cijfer CE
Vmbo-bb		0,11	0,83
Vmbo-kb	0,21	0,40	0,58
Vmbo-gt	0,48	0,57	0,29
Havo	0,71	0,75	0,34
Vwo	0,56		0,41

Tabel 5.1 Correlatie tussen afstroom en zittenblijven in oktober van het schooljaar en het gemiddelde cijfer op het centraal eindexamen in mei van het volgende schooljaar

Naar een hoger niveau

Volgens DUO (2016) vindt opstroom in de bovenbouw bijna niet plaats. Wat wel gebeurt, is stapelen. Van stapelen is sprake als een leerling na het behalen van zijn eindexamen verdergaat met onderwijs in een hogere leerweg of hoger schooltype. Figuur 5.4 bevat het percentage leerlingen dat stapelt [15]. Over de hele periode is te zien dat het percentage leerlingen dat stapelt, in vmbo-gt aanzienlijk hoger ligt dan in andere leerwegen. Wel is in vmbo-gt sprake van een duidelijke daling tot 2013, die gevolgd wordt door een stijging. In vmbo-bb en havo blijft het aantal stapelaars over de hele linie stijgen, terwijl in vmbo-kb het stapelen vrijwel constant blijft. Omdat stapelen vaak wordt gecombineerd met toelatingseisen, is het effect ervan op de eindcijfers niet te voorspellen.

15 Omdat de vermelde percentages in de verschillende Examenmonitoren niet overeenkomen worden hier alleen de cijfers uit de laatste monitor vermeld.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Rekentoets; Exameneis Nederlands aangescherpt	CE Neder-lands geijkt; (Rekentoets)	(Rekentoets)	(Rekentoets)
Havo/Vwo		Heelvak/deel-vak af-geschaft (Havo)	Heelvak/deel-vak af-geschaft (Vwo)		Aange-scherpte exameneis	Kernvakken-regeling	Rekentoets	CE Neder-lands geijkt & geleide samenvatting afgeschaft; (Rekentoets) Nieuw examenpro-gramma Natuurweten-schappen (Havo)	Rekentoets telt mee in examenresul-taat (Vwo) Nieuw examenpro-gramma Natuurweten-schappen (Vwo)	Rekentoets telt mee in kernvakken-regel (Vwo) Nieuw examenpro-gramma Wis-kunde (Havo)

Figuur 5.4 Percentage leerlingen dat stapelt

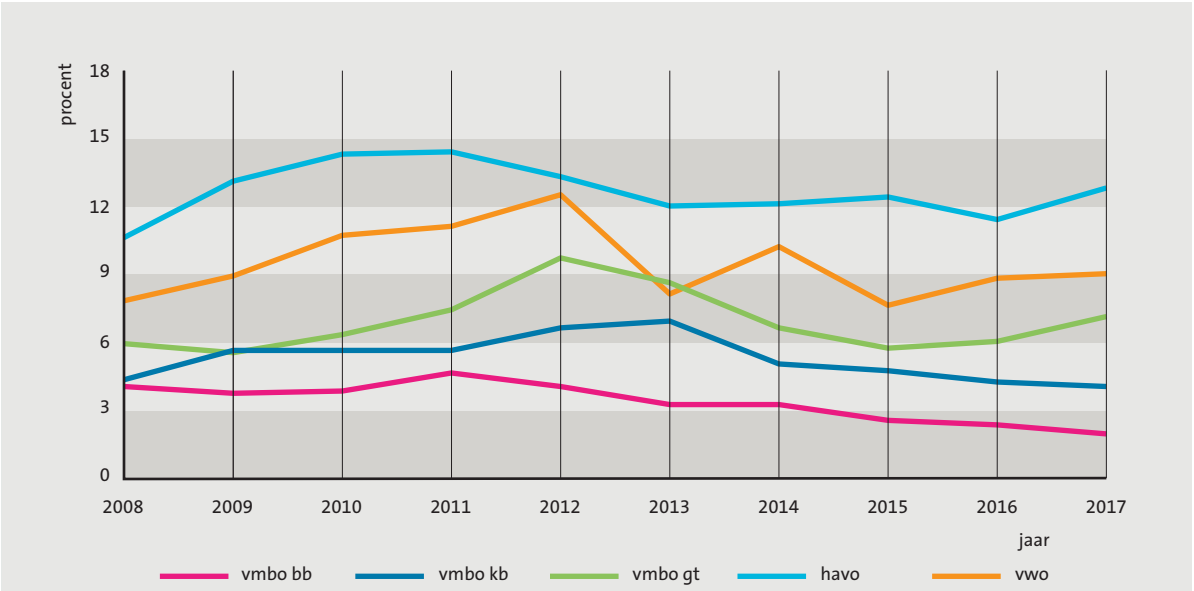
We zien dus wat wijzigingen in het stapelen, maar het effect daarvan op de eindcijfers is niet te voorspellen. Het effect van een verhoging van zittenblijven en afstroom daarentegen wel: hoe meer leerlingen zittenblijven en afstromen hoe hoger het eindcijfer een examenjaar later.
In aanloop naar de strengere exameneisen in 2012 is de doorstroom (zittenblijven en afstroom) gewijzigd. Het zittenblijven en de afstroom stijgt tot en met 2011 om daarna geleidelijk te dalen.

5.3 Geen diploma

Niet alleen de doorstroom kan wijzigen, maar ook het aantal leerlingen dat centraal eindexamen doet en slaagt.

Gezakt

Figuur 5.5 brengt het percentage gezakte leerlingen in beeld. Dat percentage neemt tot 2012 vooral toe om in de periode daarna te dalen. Het vwo laat een dip zien in 2013. Dit kan te maken hebben met een voorzichtig bevorderingsbeleid naar de examenklas. Net zoals bij het percentage zittenblijvers, is havo koploper voor het percentage gezakte leerlingen. In vmbo-bb is het percentage gezakten het kleinst.



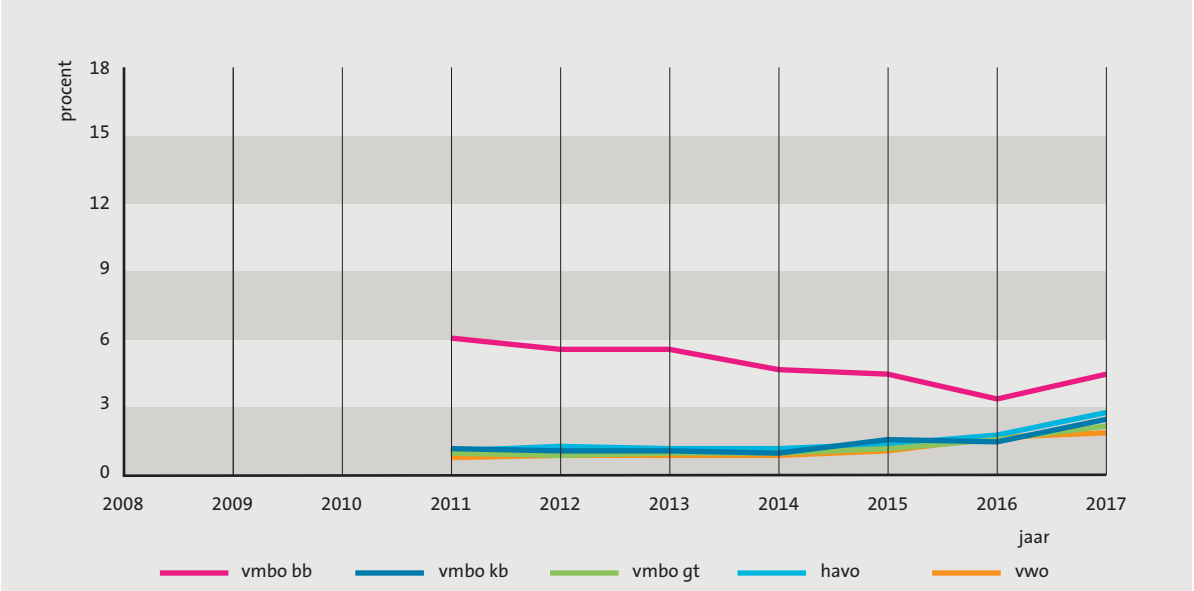
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Rekentoets; Exameneis Nederlands aangescherpt	CE Neder-lands geijkt; (Rekentoets)	(Rekentoets)	(Rekentoets)
Havo/Vwo		Heelvak/deel-vak af-geschaft (Havo)	Heelvak/deel-vak af-geschaft (Vwo)		Aange-scherpte exameneis	Kernvakken-regeling	Rekentoets	CE Neder-lands geijkt & geleide samenvatting afgeschaft; (Rekentoets) Nieuw examenpro-gramma Natuurweten-schappen (Havo)	Rekentoets telt mee in examenresul-taat (Vwo) Nieuw examen-programma Natuurweten-schappen (Vwo)	Rekentoets telt mee in kernvakken-regel (Vwo) Nieuw examen-programma Wiskunde (Havo)

Figuur 5 5 Percentage leerlingen dat is gezakt.

Niet aangemeld

Een ander fenomeen is dat niet alle leerlingen in een bepaald examenjaar ook centraal eindexamen doen. In figuur 5.6 is het percentage getoond van leerlingen uit examenklassen dat uiteindelijk niet wordt aangemeld voor het eindexamen[16]. In 2017 blijkt het percentage niet-aangemelde leerlingen in havo/vwo, vmbo-kb en vmbo-gt meer dan verdubbeld, vergeleken met 2014. In vmbo-bb ligt het percentage duidelijk hoger dan in andere schoolsoorten. In 2011 ligt het percentage niet-aangemelde leerlingen op (6%). Dat percentage is tot 2016 gedaald, om in 2017 weer te stijgen. Het percentage ligt dan nog steeds duidelijk hoger dan in de andere schoolsoorten/leerwegen (4,4%), maar het verschil is minder groot dan in 2011.
Ook het aantal niet-aangemelde leerlingen kan samenhangen met hogere cijfers op het centraal eindexamen. Uit berekeningen blijkt dat het percentage niet-aangemelde leerlingen in vmbo-kb (r = 0,42), vmbo-gt (r= 0,34), havo (r = 0,11) en vwo (r= 0,47) positief samenhangt met het gemiddeld behaalde cijfer op het centraal eindexamen. Hoe meer leerlingen in de examenklas niet zijn aangemeld voor het examen, hoe hoger het gemiddelde cijfer. Alleen in vmbo-bb, waar de trendlijn duidelijk afwijkt, is die positieve samenhang er niet (r= -0,77).
Met uitzondering van vmbo-bb zien we na het aanscherpen van de exameneis een stijging van het aantal niet-aangemelde leerlingen. Dit fenomeen hangt op zijn beurt samen met een hoger eindcijfer.

16 Omdat de vermelde percentages in de verschillende Examenmonitoren niet overeenkomen worden hier alleen de cijfers uit de laatste monitor vermeld.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Vmbo			Begin digitale afnames (kb)		Aange-scherpte exameneis		Rekentoets; Exameneis Nederlands aangescherpt	CE Neder-lands geijkt; (Rekentoets)	(Rekentoets)	(Rekentoets)
Havo/ Vwo		Heelvak/deel-vak af-geschaft (Havo)	Heelvak/deel-vak af-geschaft (Vwo)		Aange-scherpte exameneis	Kernvakken-regeling	Rekentoets	CE Neder-lands geijkt & geleide samenvatting afgeschaft; (Rekentoets) Nieuw examen-programma Natuurweten-schappen (Havo)	Rekentoets telt mee in examen-resul-taat (Vwo) Nieuw examen-programma Natuurweten-schappen (Vwo)	Rekentoets telt mee in kernvakken-regel (Vwo) Nieuw examen-programma Wiskunde (Havo)

Figuur 5.6. Percentage leerlingen in de examenklas dat niet is aangemeld voor het examen

5.4 Samenvatting en conclusie

Eén factor die van invloed kan zijn op het gemeten prestatieniveau bij centrale examens en in PISA is de leerlingpopulatie. De leerlingpopulatie kan tussen de PISA-afname bij een representatieve steekproef en het eindexamen nog wijzigen. In dit hoofdstuk is bekeken of we veranderingen zien in de verdeling van leerlingen over de onderwijsniveaus, hun doorstroom en het verkrijgen van een diploma in het eindexamenjaar. We zien ook een relatie tussen dergelijke veranderingen en veranderingen in het gemiddelde cijfer op het centraal eindexamen.

De doorstroom van leerlingen over de afgelopen tien jaar is niet constant. De doorstroom fluctueert licht, met een zichtbaar keerpunt in 2011. Alhoewel de verdeling van leerlingen over schooltypen/leerwegen in de bovenbouw vrij constant blijft, zien we vanaf 2012 een afname van het aantal leerlingen dat afstroomt naar een lager niveau of blijft zitten. Een afname in afstroom of zittenblijven hangen over het algemeen samen met een daling van het gemiddelde examencijfer in het daarop volgende schooljaar. Dit zou dus een verklaring kunnen vormen voor de waargenomen afgevlakte stijging in 2014 en later. Tegelijkertijd moeten we constateren dat het percentage leerlingen in het examenjaar dat niet aangemeld wordt voor het centraal examen, stijgt vanaf 2011 (met uitzondering van vmbo-bb). Dit hangt over het algemeen juist weer samen met een stijging van het gemiddelde examencijfer.

Het keerpunt 2011 in de stijgende trend van zittenblijven en afstroom is opvallend. In hoofdstuk 2 zagen we ook bij de gemiddelde cijfers een keerpunt in 2011. Met de aanscherping van de exameneis in 2012 begon het gemiddelde cijfer op het centraal eindexamen te stijgen.
De aangescherpte maatregelen lijken dus zowel een invloed te hebben gehad op de CE-cijfers, als op de doorstroom. De doorstroom op zijn beurt heeft een samenhang met de eindexamencijfers een schooljaar later.

6 De rol van motivatie bij toetsresultaten

6 De rol van motivatie bij toetsresultaten

Toetsen worden afgenomen om uitspraken te kunnen doen over het vaardigheidsniveau van leerlingen of groepen leerlingen. De prestaties op de toetsen zouden uiteraard bepaald moeten worden door de vaardigheid van de leerlingen waarin men geïnteresseerd is. Maar toetsresultaten kunnen ook beïnvloed worden door niet-cognitieve factoren. Een van deze niet-cognitieve factoren is de motivatie van leerlingen om de toets te maken.

In het eerste deel van dit hoofdstuk wordt nader ingegaan op de rol van motivatie bij de PISA-resultaten in Nederland. PISA is een onderwijskundig survey waarin de toetsresultaten van een steekproef van individuele leerlingen worden gebruikt om uitspraken te doen over het vaardigheidsniveau van alle 15-jarige leerlingen in een land. Deze resultaten kunnen dan gebruikt worden om de uitkomsten gevonden in een land cross-sectioneel te vergelijken met andere landen én met eerdere PISA-resultaten binnen het land zelf (longitudinale trendvergelijkingen).

Op het moment dat de motivatie van leerlingen om deel te nemen aan het PISA-onderzoek verschillend is tussen landen, kan dat de cross-sectionele vergelijkingen vertekenen. Op het moment dat de motivatie om deel te nemen aan het onderzoek in vergelijking met eerdere afnames binnen een land is veranderd, kan dat effect hebben op de interpreteerbaarheid van de trendresultaten. Binnen dit hoofdstuk proberen we beide vragen voor Nederland te adresseren.

In deel twee van dit hoofdstuk gaan we in op het belang van examenresultaten. De invoering van de kernvakkenregeling op havo en vwo in 2013 was een belangrijke wijziging in de zak/slaagregeling. Een leerling met een vier op zijn eindlijst voor een kernvak kreeg vanaf 2013 geen diploma meer. De invoering van de kernvakkenregel deed daarmee het belang van de kernvakken groeien. In paragraaf 6.2 onderzoeken we de gevolgen van de invoering van de kernvakkenregel en in paragraaf 6.3 onderzoeken we de gevolgen van de invoering van de rekentoets als voorbeelden van een grote verandering in relatie tot het resultaat. We beschrijven daarbij wat de gevolgen in deze situaties waren en wat in algemene zin de gevolgen kunnen zijn wanneer het belang van de toets verandert.

6.1 Het effect van motivatie op toetsresultaten

Verschillen in de motivatie van leerlingen om een toets af te nemen, kunnen onder andere verklaard worden door verschillen in toetscondities. Een veelgemaakt onderscheid hierbij is het verschil tussen *low*- en *high-stakes* toetscondities (Holland & Wightman, 1982). In een low-stakes toetsconditie zijn er geen consequenties verbonden aan de toetsprestatie. Omgekeerd geldt dat hoe meer high-stakes de toetscondities zijn, hoe groter de persoonlijke gevolgen voor leerlingen zijn.

De verwachting is dat leerlingen onder high-stakes toetscondities harder werken en streven naar een maximale prestatie, terwijl low-stakes toetscondities juist een typische prestatie bij leerlingen oproept (Keizer-Mittelhaëuser, 2014). Een gevolg daarvan kan zijn dat er verschillen ontstaan in de

moeite die leerlingen nemen om de toets voor te bereiden, vragen binnen de toets te beantwoorden en daarmee ook de prestatie die zij uiteindelijk leveren (Wise & DeMars, 2005).

Internationale surveys zoals PISA zijn typische low-stakes toetsen. Studenten bereiden zich niet voor op de toets, krijgen geen feedback over hun prestatie en er zijn geen persoonlijke gevolgen verbonden aan de toetsprestaties. Meerdere studies hebben laten zien dat leerlingen niet maximaal presteren op het moment dat er geen feedback gegeven wordt of als de prestatie op een vraag niet mee telt bij de bepaling van een toetsscores (zie onder andere Wise & DeMars, 2005; O'Neill, Sugrue, & Baker, 1996; Kiplinger & Linn, 1996, cf. Keizer-Mittelhaëuser, 2014).

Wise & DeMars (2005) lieten in een meta-analyse zien dat het verschil in prestaties tussen gemotiveerde en ongemotiveerde leerlingen een gemiddelde *effect size* van 0.59 heeft. Op de PISA-schaal komt dit overeen met ruim 50 scorepunten. Wise & Kong (2005) suggereren dat een lage motivatie om een toets te maken niet per se constant hoeft te zijn. Het zou kunnen zijn dat leerlingen binnen een toets meer of minder aandacht aan opgaven besteden naarmate de toets vordert. Dit wordt verder ondersteund door een studie van Wise (2007) waarin wordt beargumenteerd dat leerlingen weliswaar gemotiveerd kunnen beginnen aan een toets, maar op een gegeven moment steeds minder gemotiveerd kunnen worden.

Een studie van Wolf, Smith en Birnbaum (1995) laat zien dat ook het type item het effect van motivatie op toetsresultaten kan beïnvloeden. De prestaties op items die een hogere cognitieve last van de leerlingen vragen, zouden meer beïnvloed worden door een (gebrek aan) motivatie dan vragen die sneller te beantwoorden zijn.

Een recentere studie van Gneezy, List, Livingston, Sadoff, Qin & Xu (2017) probeert door middel van het variëren van externe beloningen in een experimentele setting vast te stellen wat het effect van (interne) motivatie op toetsresultaten is. Door een groep leerlingen een beloning in het vooruitzicht te stellen die afhankelijk is van de toetsprestatie, en deze prestatie te vergelijken met leerlingen in de controlegroep, hebben de onderzoekers geprobeerd het effect van motivatie vast te stellen[17]. Het onderzoek liet zien dat leerlingen in Shanghai mét een beloning niet beter presteerden dan leerlingen zónder beloning. In de Verenigde Staten bleek er wel een groot verschil te zijn (een effect size van ongeveer 0.2, dit komt ongeveer overeen met 20 scorepunten op de PISA-schaal). Dit wijst uit dat leerlingen in de Verenigde Staten in low-stakes toetsen niet optimaal gemotiveerd zijn. Dit in tegenstelling tot leerlingen in Shanghai (Gneezy et al., 2017). Het onderzoek sluit aan bij eerdere studies die ook hogere proxy-waarden voor motivatie rapporteerden in Oost-Aziatische landen (Zamarro, Hitt & Mendez, 2016, cf. Gneezy et al., 2017).

6.1.1 Data en operationalisatie

Binnen dit onderzoek maken we gebruik van data uit PISA-onderzoek. Voor het evalueren van verschillen in motivatie tussen landen wordt gebruik gemaakt van PISA-2015. Hierbij worden de uitkomsten in Nederland vergeleken met andere landen in de Europese Unie (EU) en met Oost-Aziatische landen[18]. We hebben voor deze laatste vergelijking gekozen omdat Oost-Aziatische landen hoog scoren en omdat er zoals in de vorige paragraaf genoemd was een hoge intrinsieke motivatie binnen deze landen is gevonden.

17 In de studie kregen leerlingen 25 vragen voorgelegd. Leerlingen in de experimentele conditie kregen 1 US dollar voor elke vraag die zij goed hadden beantwoord. Studenten in de controlegroep kregen geen beloning (Gneezy et al, 2017).

18 Binnen dit onderzoek zijn de volgende landen/entiteiten tot Oost-Azië gerekend: Hong Kong, Japan, Macao, China, Singapore en Taiwan.

In 2015 hebben in Nederland 5385 leerlingen aan PISA deelgenomen. In de EU exclusief Nederland zijn er 156 990 participanten en binnen de Oost-Aziatische landen zijn dit er 40 146. Voor het evalueren van verschillen in motivatie tussen afnames wordt gebruik gemaakt van PISA-2012 en PISA-2015 gegevens binnen Nederland. In 2012 hebben 4 460 leerlingen in Nederland deelgenomen aan het onderzoek. De PISA-onderzoeken zijn zodanig opgesteld dat de uiteindelijke resultaten representatief zijn voor alle 15-jarige leerlingen in een land.

Tot en met de PISA-afname in 2012 is (grotendeels) gebruik gemaakt van een papieren afname van de toetsen. In 2015 is de PISA-studie voor de eerste keer grootschalig afgenomen met behulp van een computer. Leerlingen maakten in 2015 in totaal vier clusters met opgaven: twee clusters in het eerste uur, en na een pauze, twee clusters in nog een uur.

Helaas is er geen directe meting van motivatie binnen PISA beschikbaar en daardoor maken we gebruik van proxy-metingen van motivatie. Dit zijn achtereenvolgens een in de PISA achtergrondvragenlijst opgenomen *prestatiemotivatie-index*, *responstijden* en het percentage door leerlingen overgeslagen vragen binnen een toets (*item nonrespons*). De eerste twee gegevens gebruiken we om verschillen tussen landen te onderzoeken. Item nonrespons om verschillen tussen PISA-2012 en 2015 binnen Nederland te evalueren.

Prestatiemotivatie-index
In PISA-2015 is voor de eerste keer een zogenaamde prestatiemotivatie-index in de leerlingenachtergrondvragenlijst opgenomen (OECD, 2017). Hogere waarden op deze index geven aan dat de leerlingen een hogere prestatiemotivatie hebben.

Responstijden
Het digitale platform, gebruikt in 2015, slaat ook de hoeveelheid tijd op die leerlingen besteden aan elke vraag in de toets (OECD, 2017). Deze responstijden zijn in milliseconden geregistreerd. Deze informatie wordt als proxy indicator gebruikt voor motivatie. De hypothese is dan dat een korte responstijd samen met een fout antwoord, duidt op een lagere motivatie.

Voordat de analyses zijn uitgevoerd, zijn eerst leerlingen met extreme responstijdwaarden uit de dataset verwijderd. Leerlingen die minder dan 100 000 of meer dan 20 000 000 milliseconden aan de opgaven per onderdeel hebben besteed, zijn niet meegenomen in de analyses. Dit komt overeen met respectievelijk ongeveer 2 minuten en 6 uur toetstijd. Daarmee viel een klein aantal leerlingen af (0,12%).

De responstijden worden in de resultatensectie gepresenteerd in minuten en als centrummaat is de mediaan gebruikt[19]. Eerst vergelijken we de responstijden binnen de verschillende deelnemende landen en vervolgens relateren we de individuele responstijden aan verschillende kenmerken van de items en scores op de items. Aansluitend bij de bevindingen in de literatuur, worden de relaties tussen responstijd en 1) positie in de toets, 2) itemtype en 3) herkomstland van de leerlingen geëvalueerd. Hiervoor worden de gegevens van de reguliere toetsversies wiskunde gebruikt (toetsversies 43-54), die telkens onder andere bestaat uit twee clusters met wiskundeitems[20]. In totaal zijn er zes clusters met verschillende opgaven wiskunde. Het toetsafnamedesign is zodanig opgesteld dat ieder cluster op iedere positie in de toets is afgenomen (OECD, 2017).

19 Er is voor de mediaan gekozen omdat deze centrummaat minder gevoelig is voor extreme waarden dan bijvoorbeeld het gemiddelde.

20 Vanwege technische redenen en om het aantal gepresenteerde resultaten te beperken, analyseren we alleen de wiskundeopgaven in de reguliere opgavenboekjes.

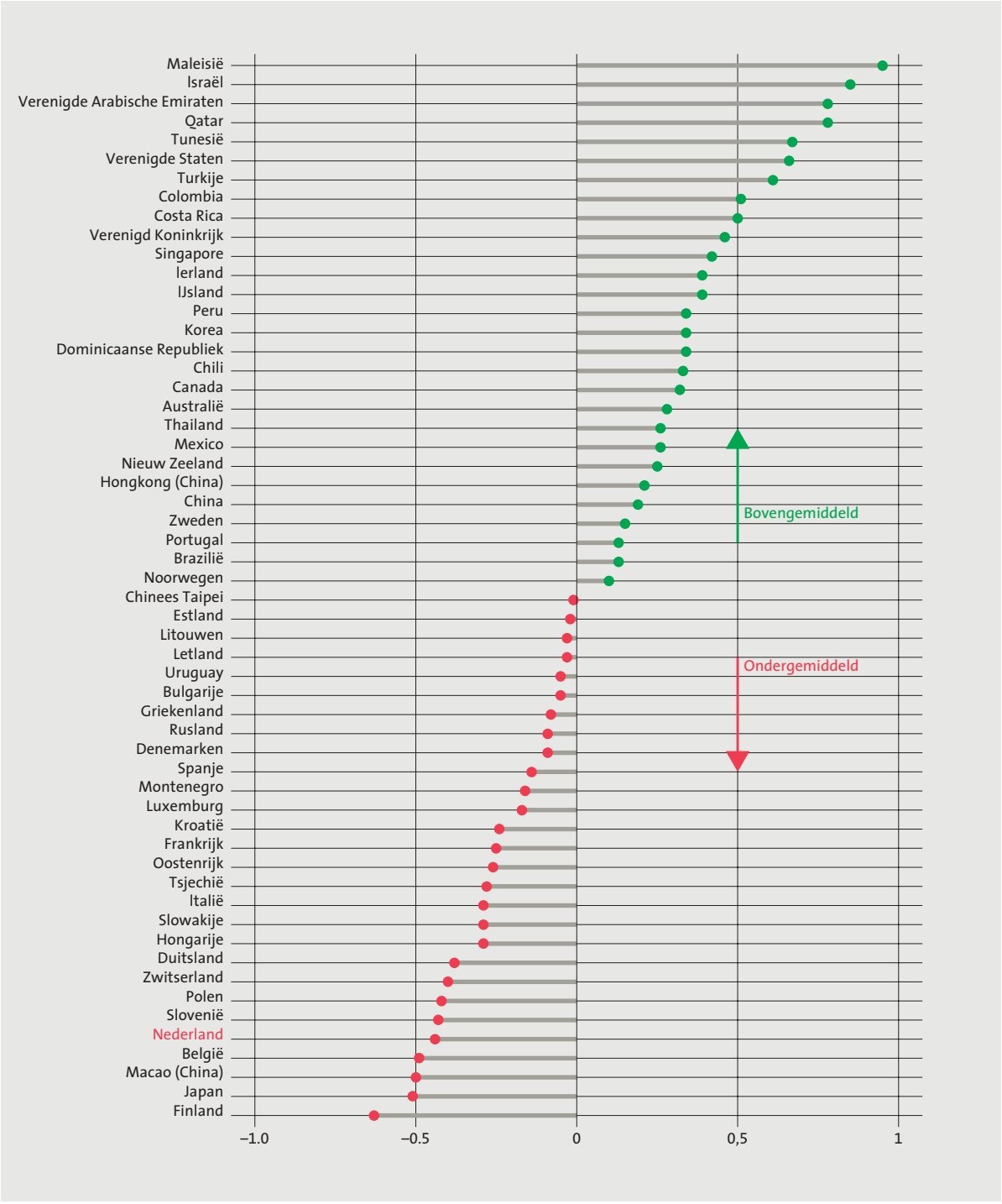
Item nonresponse
Binnen PISA-2012 is een onderscheid gemaakt tussen vier verschillende types ontbrekende data voor de cognitieve toetsen: 1) Item nonresponse, in dit geval is een antwoord op een vraag verwacht, maar is er geen antwoord op de vraag gegeven door de leerling; 2) meerdere of niet-geldige antwoorden, een categorie die aangeeft dat een leerling meerdere antwoorden op een vraag heeft geselecteerd, terwijl maar één antwoord verwacht was; 3) 'system missing', item nonresponse die betrekking heeft op vragen die niet voorkwamen in de toetsversie die de leerling heeft gemaakt of naar vragen die na de toetsafname zijn uitgesloten van verdere analyses door bijvoorbeeld vertaalfouten in de vraag zelf; 4) vragen aan het eind van de toets die niet meer beantwoord zijn door de leerling (OECD, 2014).

In 2015 is een vijfde type ontbrekende data categorie toegevoegd. Dit is de categorie "niet van toepassing". Dit zijn vragen die leerlingen hadden moeten overslaan, maar toch hebben beantwoord. Bovendien is de derde categorie ('system missing') veranderd. In 2015 heeft deze categorie niet alleen betrekking op vragen die niet in aan de leerling zijn voorgelegd, maar ook op die gevallen waar de toetsafname eerder dan verwacht stopte (OECD, 2017).

6.1.2 Resultaten

Uitkomsten op de prestatiemotivatie-index

In figuur 6.1 zijn de landengemiddelden op de prestatiemotivatie-index gepresenteerd.



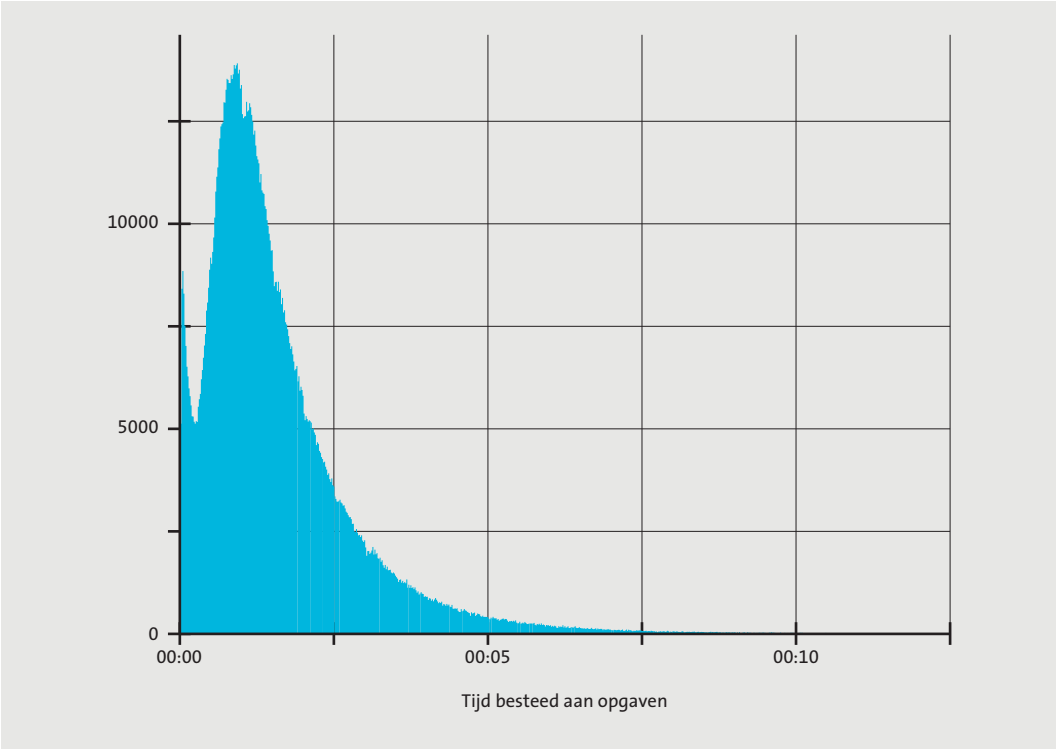
Figuur 6.1 Prestatiemotivatie-index waarden

In vergelijking met de andere EU-landen scoort Nederland niet hoog (-0.44). Alleen Finland (-0.63) en België (-0.49) scoren nog lager. Hoog scorende EU-landen zijn vooral Groot-Brittannië (0.46) en Ierland (0.44). Wel is opvallend dat juist de relatief hoog scorende landen, relatief laag

scoren op de cognitieve domeinen. In hoeverre de studentenantwoorden op vragen in de index tussen landen goed onderling vergelijkbaar zijn, kan dan ook in twijfel worden getrokken.

Responstijden

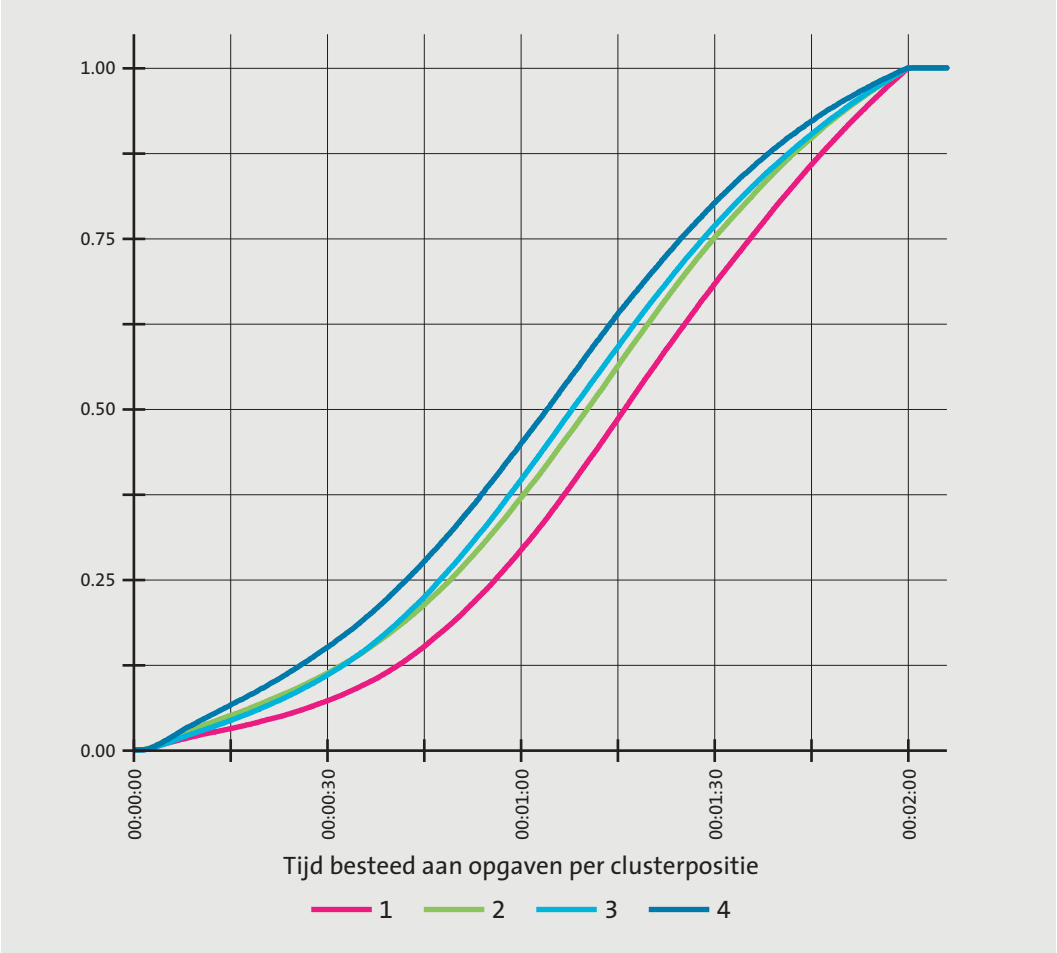
Figuur 6.2 geeft een beeld van de tijd die leerlingen besteden aan een opgave. Twee verdelingen zijn zichtbaar: leerlingen die zeer weinig tijd aan een opgave besteden en leerlingen die meer tijd aan opgaven besteden. Veruit de meeste leerlingen besteden gemiddeld niet meer dan 5 minuten aan een opgave. De mediaan is 1 minuut en 15 seconden.



Figuur 6.2 Tijd besteed aan opgaven

Itempositie

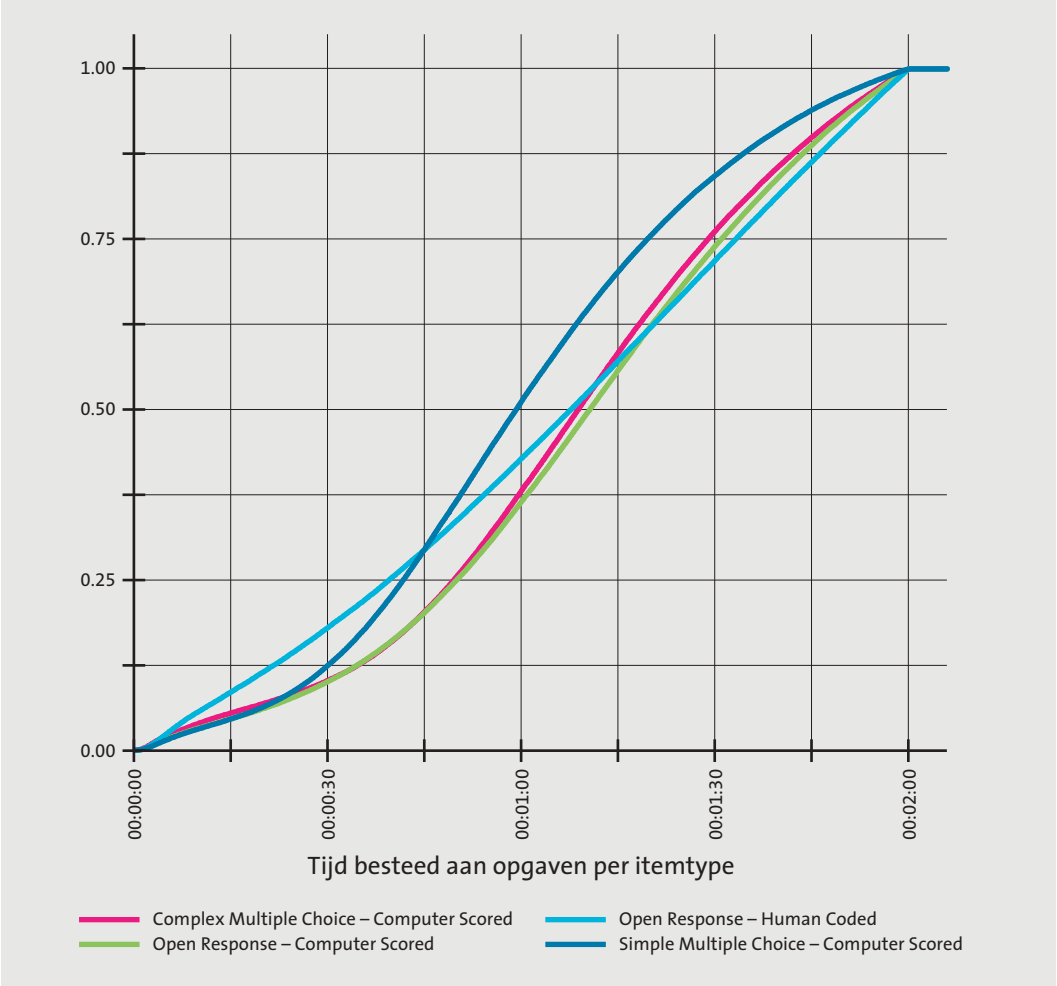
Een van de hypothesen uit onderzoek is dat de motivatie tijdens het maken van een toets, daalt. Deze hypothese onderzoeken we in figuur 6.3. Daaruit blijkt inderdaad duidelijk dat de tijd die leerlingen besteden aan opgaven, daalt naarmate de toets vordert. Waar de helft van de leerlingen minstens 1 minuut 15 besteedt aan opgaven in het eerste cluster (de rode verdelingsfunctie), wordt aan dezelfde opgaven in cluster 4 (de paarse verdelingsfunctie) nog ruim 1 minuut besteed. Dat is relatief gezien een behoorlijke daling.



Figuur 6.3 Tijd besteed aan opgaven per cluster

Itemtype

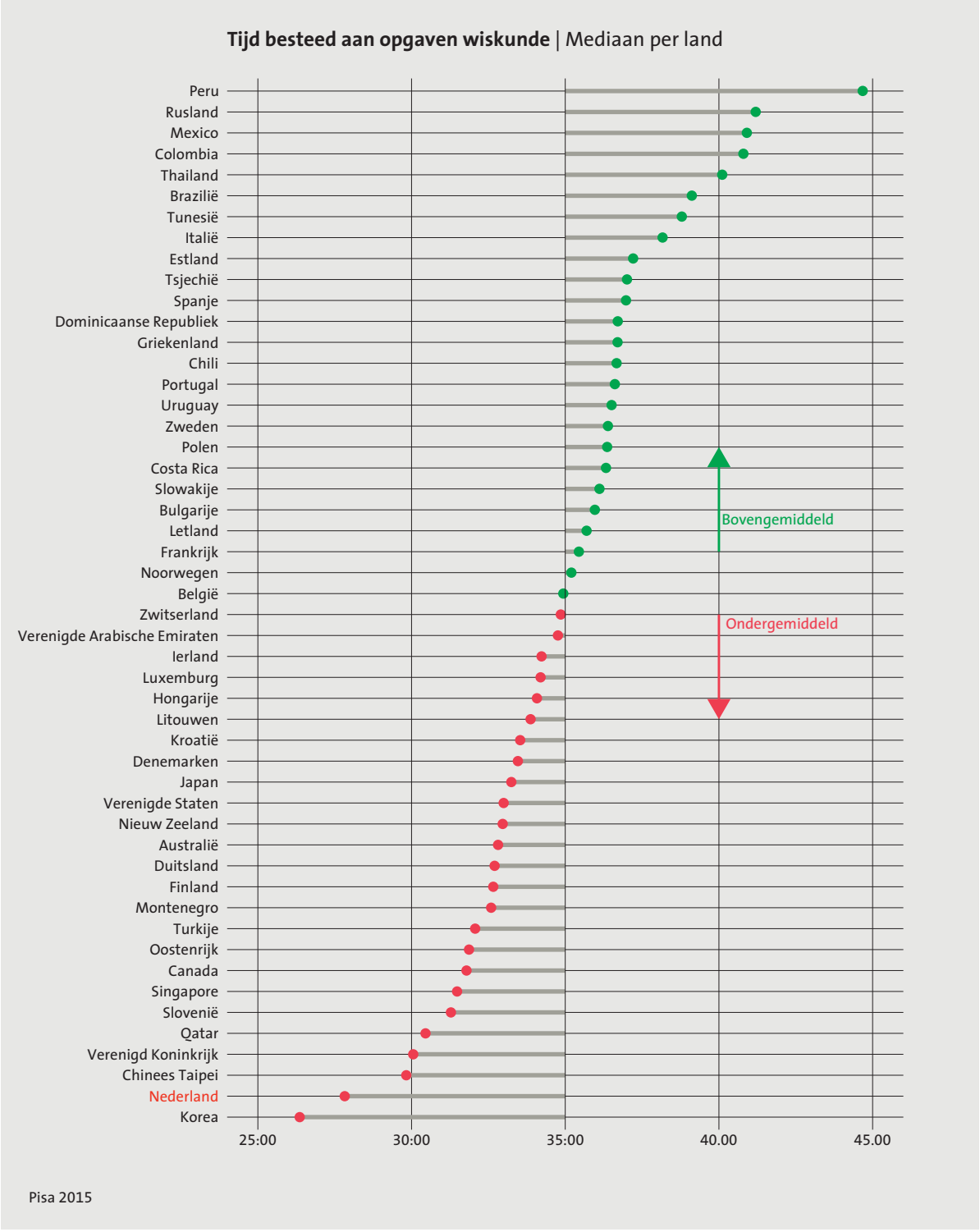
PISA maakt gebruik van verschillende vraagtypen. Meerkeuze-items komen het meest voor, maar de toets bevat ook open responsvragen (vragen waarbij leerlingen zelf het antwoord moeten formuleren). Eenvoudige open responsvragen worden binnen PISA met behulp van een computer nagekeken. Meer complexere open responsvragen worden met behulp van een beoordelingsvoorschrift door beoordelaars gescoord. Een verschil in responstijd op de verschillende vraagtypen kan wijzen op een verschil in motivatie. Figuur 6.4 presenteert de responstijden per itemtype. Niet onverwacht besteden leerlingen de minste tijd aan eenvoudige meerkeuze-items (de paarse verdelingsfunctie). Opvallend is dat complexere open respons-vragen van de ene groep leerlingen relatief weinig tijd vragen, en van een andere groep juist relatief veel (de turquoise gekleurde verdeling).



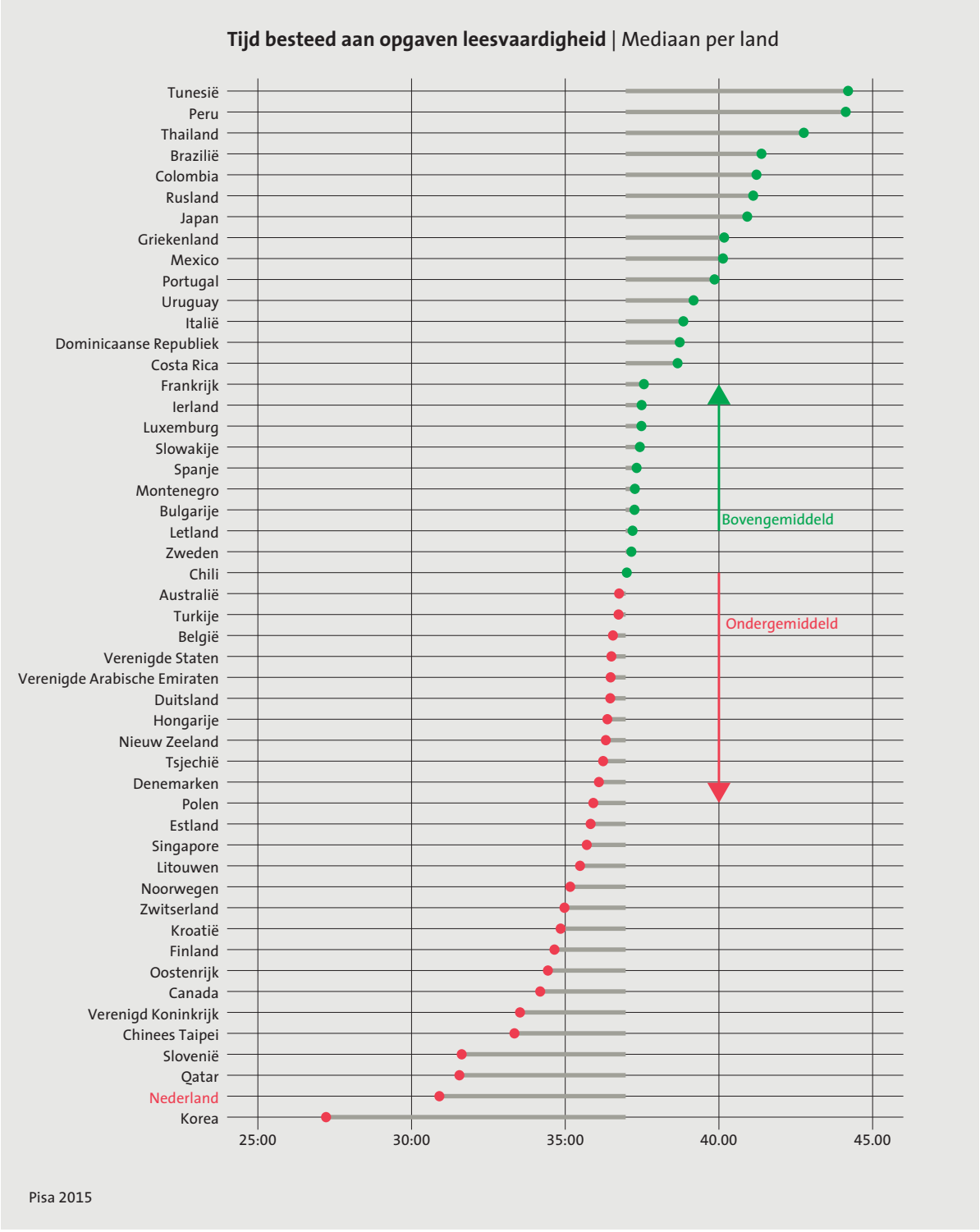
Figuur 6.4 Tijd besteed aan opgaven per itemtype

Herkomstland van de leerlingen

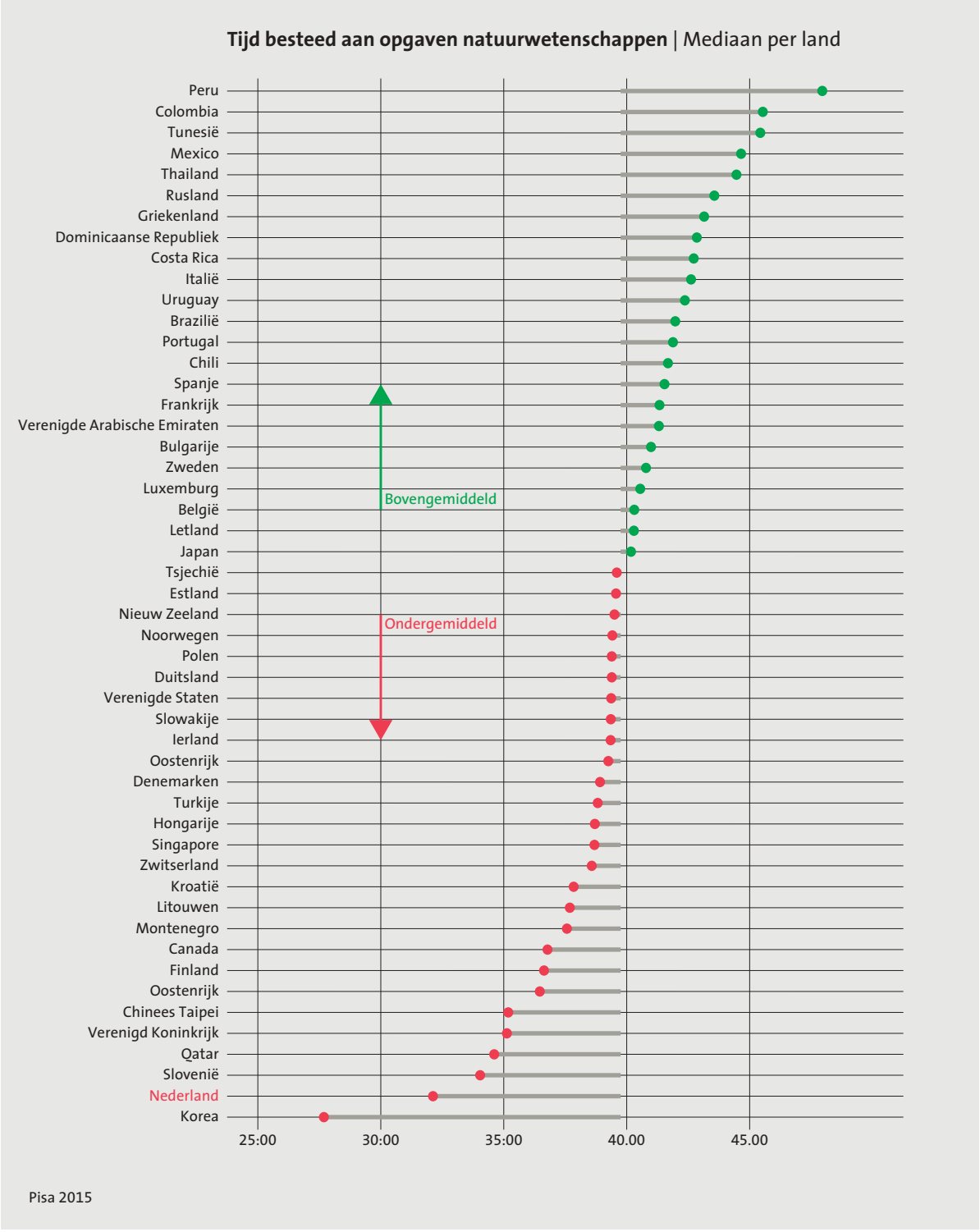
Wat zijn de verschillen in tijd die Nederlandse leerlingen aan een item besteden, vergeleken met de tijd die leerlingen uit andere landen uit de EU aan het item besteden? Figuren 6.5, 6.6 en 6.7 maken dit duidelijk voor respectievelijk de toetsen natuurwetenschappen, leesvaardigheid en wiskunde. Van alle landen in de EU besteden leerlingen in Nederland de minste tijd aan de toets. Dit geldt zowel voor natuurwetenschappen, leesvaardigheid en wiskunde. Waar in Nederland de mediaan van het aantal besteedde minuten aan de leesvaardigheidstoets 30:54 is, ligt dat in België (36:33) en Duitsland (36:27) een stuk hoger. In de Zuid-Europese landen wordt binnen de EU de meeste tijd aan de toetsafname besteed. Zuid-Korea is het enige land waar leerlingen minder tijd aan de opgaven besteden dan in Nederland.



Figuur 6.5 Responstijden (mediaan) voor wiskunde

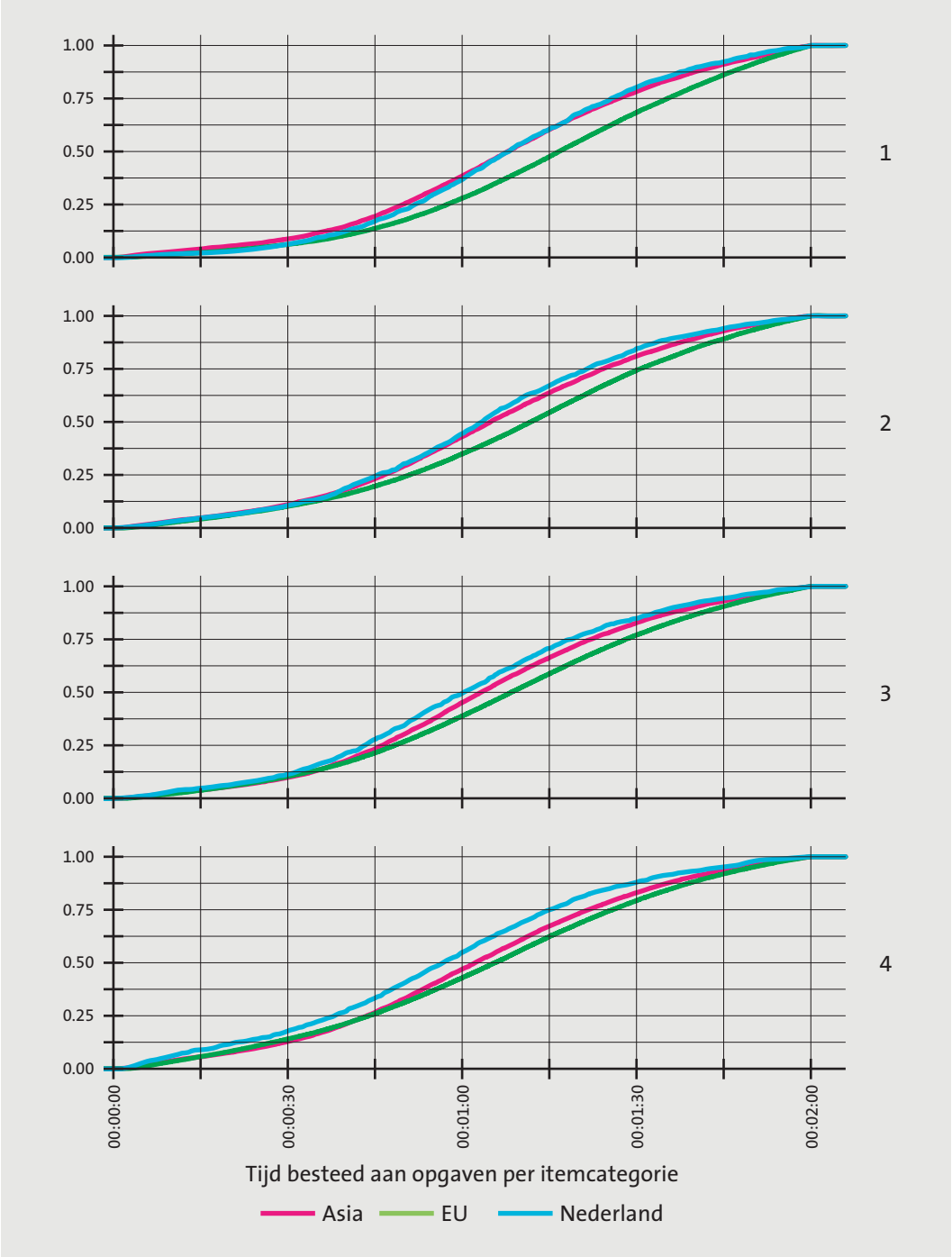


Figuur 6.6 Responstijden (mediaan) voor leesvaardigheid



Figuur 6.7 Responstijden (mediaan) voor natuurwetenschappen

De figuur 6.8 relateert de hoeveelheid tijd die aan opgaven wordt besteed afhankelijk van de positie van de positie van het cluster in de toets, maar nu uitgesplitst naar de groepen Nederlandse-, EU- en Oost-Aziatische leerlingen.

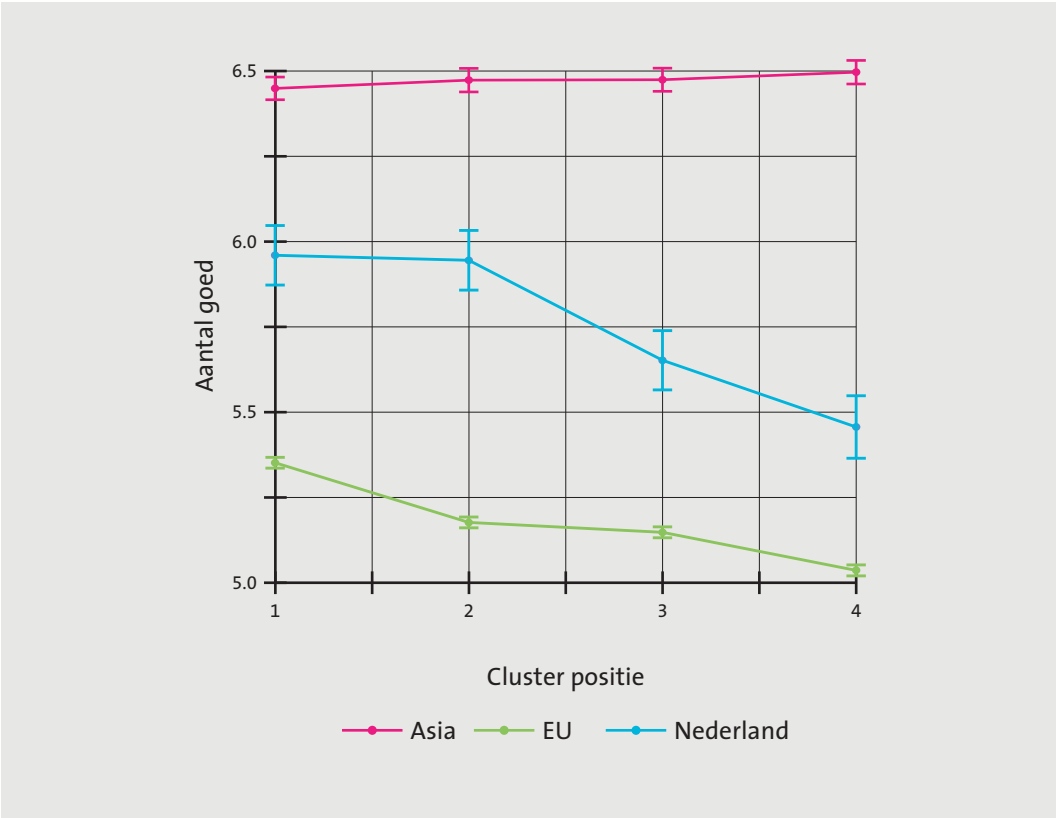


Figuur 6.8 Tijd besteed aan opgaven per clusterpositie per herkomstland

Over de hele linie besteden leerlingen minder tijd aan het vierde cluster. Dit verschil is relatief klein in Oost-Azië, zeker in vergelijking met de EU en Nederland. Het verschil in tijd dat leerlingen in Nederland en Oost-Azië aan opgaven besteden, wordt groter naarmate de toets vordert. Terwijl de responstijd in cluster 1 nog vergelijkbaar is (de verdelingsfuncties overlappen vrijwel), is duidelijk te zien dat dit in cluster 4 niet meer geldt.

Omdat dezelfde opgaven in alle vier de clusters voorkomen, is het interessant om het aantal goed gemaakte opgaven per cluster te vergelijken. Deze resultaten, die te vinden zijn in figuur 6.9, zijn opmerkelijk. Terwijl het gemiddeld aantal juist beantwoorde opgaven in

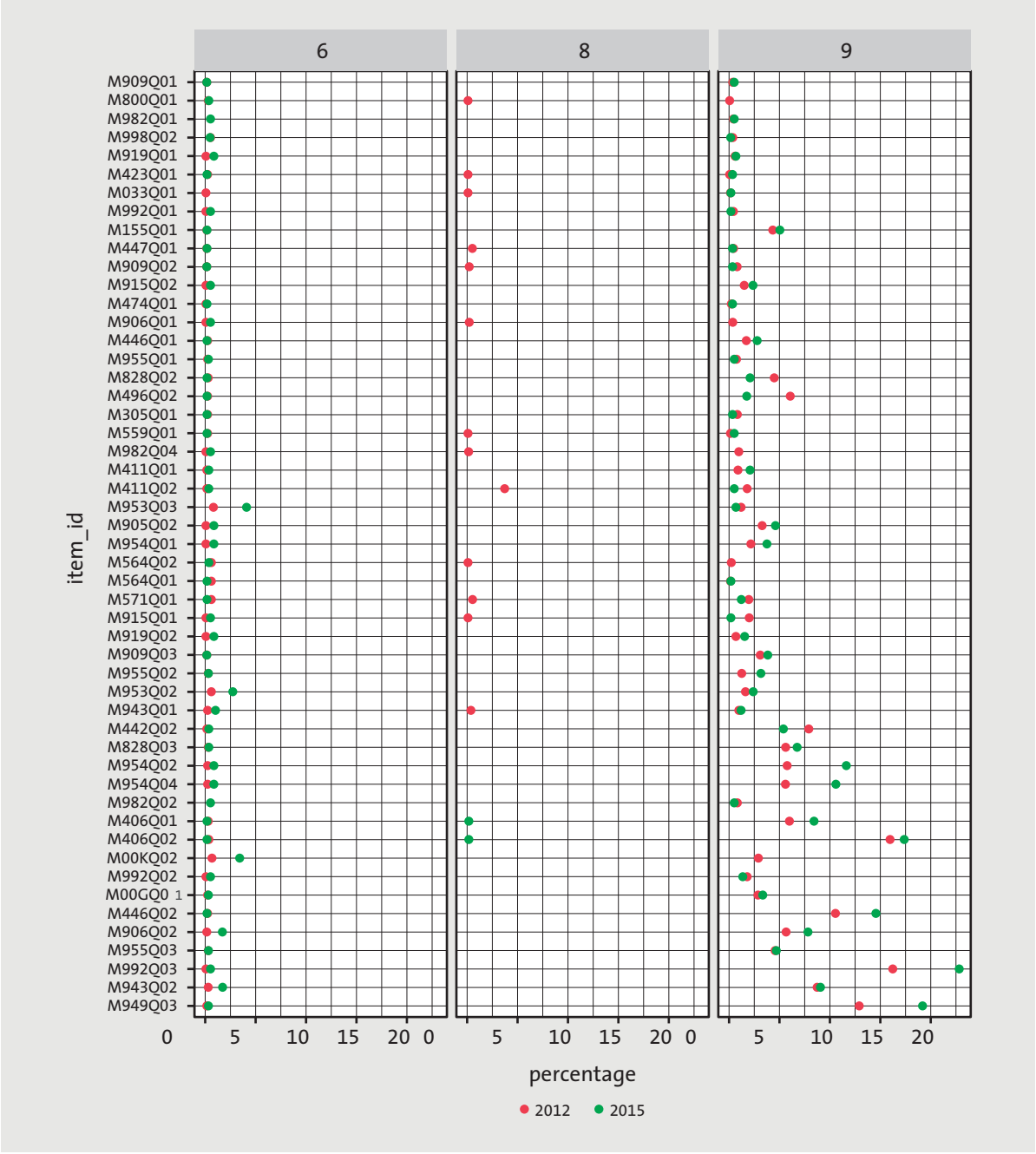
Oost-Azië constant blijft, daalt het gemiddelde in Nederland en de EU. Deze daling is in Nederland groter dan in de rest van de EU.



Figuur 6.9 Gemiddeld aantal opgaven goed per clusterpositie per herkomstland

Item nonresponse in PISA-2012 en 2015

In figuur 6.10 zijn de percentages ontbrekende antwoorden gepresenteerd voor items die zowel in wiskunde 2012 en 2015 zijn afgenomen. Hierbij staan de categorieën 6, 8 en 9 voor respectievelijk “het percentage niet behaalde items”, “het percentage niet-geldige antwoorden” en het “percentage overgeslagen vragen”.



Figuur 6.10 Item nonresponse in 2012 en 2015

Voor veruit de meeste vragen geldt dat het percentage niet-bereikte vragen (categorie 6) in 2015 niet wezenlijk veranderd is ten opzichte van 2012. De vragen die in 2015 wel een hoger percentage “niet bereikt” hebben, scoren op de andere item nonresponse-categorieën een hoger percentage dan in 2012, zoals bijvoorbeeld item M953Q02. Uit een analyse van alle vragen blijkt dat het aantal overgeslagen vragen in 2012 en 2015 niet significant verschillend is. Dit geldt ook als we alleen kijken naar de vragen die in het laatste kwart voorkwamen van de toets die de leerlingen gemaakt hadden.

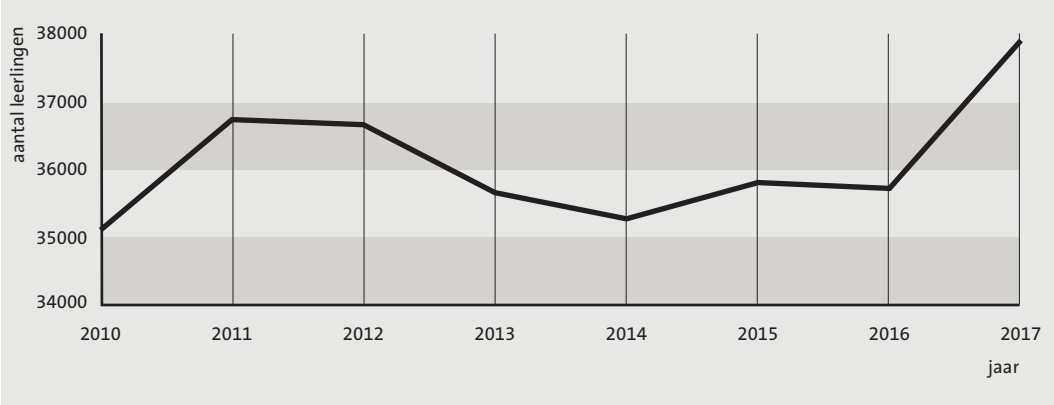
6.2 Belang van de centrale examens

Tegenover een low stakes-toets als PISA staan de centrale examens: in Nederland high stakes-toetsen bij uitstek. Diploma's vormen voor leerlingen hét toegangsbewijs voor het vervolg-onderwijs, de basis voor verdere stappen in je leven. Als het belang van een toetsresultaat toeneemt, neemt ook de aandacht van leerlingen en scholen voor die toets toe. Hoe? Dat maken we duidelijk aan de hand van de veranderde status van het vak wiskunde.

In dit onderzoek bekijken we de prestaties van vwo-leerlingen op het vak wiskunde over de afgelopen tien jaar. We maken daarbij gebruik van de gegevens in BRON van DUO. Hiermee kan ook de impact van de herkansing goed in beeld worden gebracht. We kiezen voor wiskunde, omdat dit vak bij de invoering van de nieuwe zak/slaagregeling de grootste angel bleek. De impact van de nieuwe regels was hier voor scholen en leerlingen het grootst. Als gevolg van de kernvakkenregeling moesten leerlingen voor wiskunde minimaal een 5 op hun eindlijst hebben om te slagen. Dat betekende dat in 2010 en 2011 gemiddeld 4,3% van de leerlingen zou zijn gezakt als gevolg van de kernvakkenregel. De impact van de andere kernvakken was aanzienlijk kleiner. Vanwege een 4 voor Engels zou om die reden 0,9% het diploma niet hebben gehaald en voor Nederlands 0,2%. Dat we in het onderzoek focussen op het vwo, komt omdat alle leerlingen op dit niveau eindexamen wiskunde doen. Daarbij is een keuze mogelijk tussen wiskunde A, wiskunde B of wiskunde C.

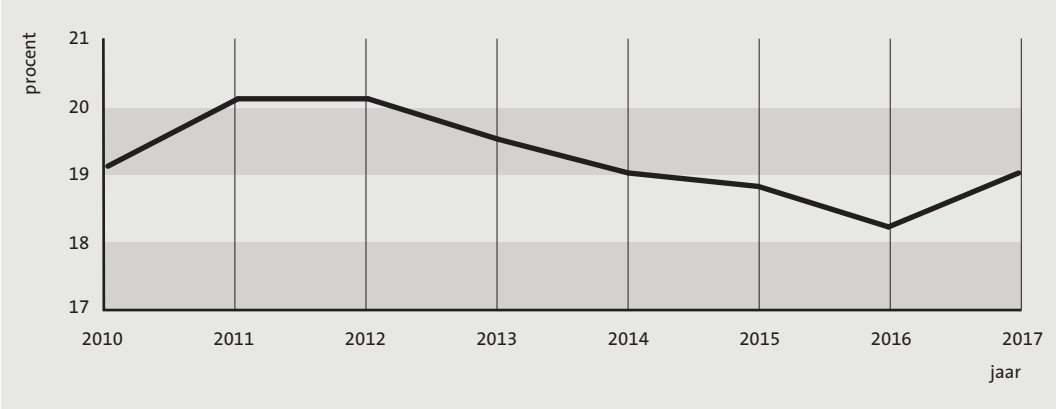
6.2.1 Examenkandidaten vwo

De afgelopen 10 jaar is het aantal vwo-examenkandidaten gegroeid. Uit figuur 6.13 blijkt dat die groei allesbehalve constant was. In de periode 2012-2014 vond een opmerkelijke daling plaats van het aantal leerlingen. Vermoedelijk had dit te maken met het plaatsingsbeleid van scholen als gevolg van de introductie van de nieuwe zak/slaagregeling.



Figuur 6.13 Absolute aantallen examenkandidaten vwo (BRON, DUO)

Figuur 6.14 geeft het aantal vwo-leerlingen weer als percentage van het totaal aantal examenkandidaten. Dit geeft een iets ander beeld. Dit wordt veroorzaakt door de gestage groei van het totaal aantal examenkandidaten van 183.000 in 2010 naar 199.000 in 2017.

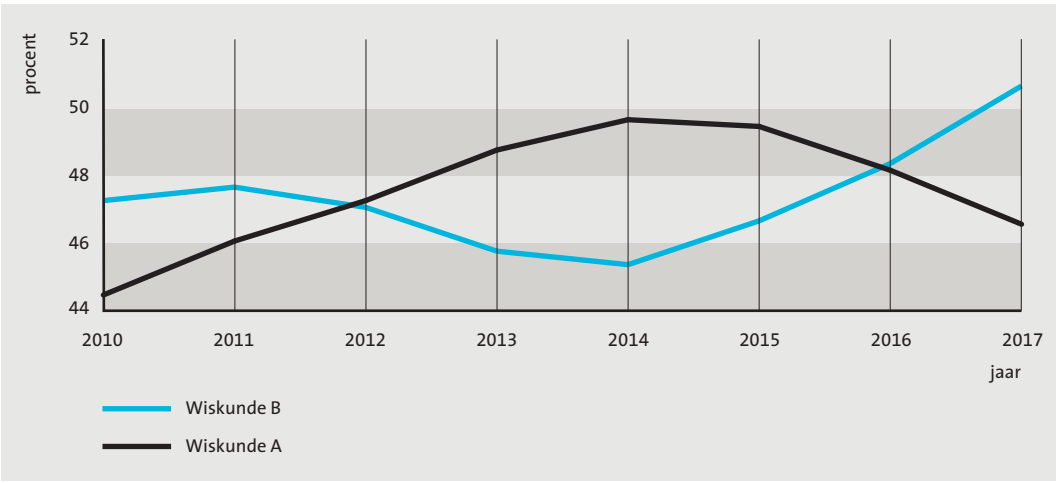


Figuur 6.14 Percentage vwo-examenkandidaten ten opzichte van alle examenkandidaten (BRON, DUO)

6.2.2 Examenkandidaten vwo-wiskunde

Vóór 2010 was het vak wiskunde opgesplitst in vier typen: B1, B1,2, A1 en A1,2. In 2010 ging de verdeling over naar de vakken wiskunde A, B en C. Figuur 6.15 laat zien hoe het aantal vwo-kandidaten met wiskunde A en B zich sindsdien ontwikkelde.

In 2014 was het aandeel vwo-leerlingen dat koos voor wiskunde B op zijn laagst. Mogelijk had dit te maken met een voorzichtige keuze van scholen en leerlingen om in de aanloop naar de introductie van de kernvakkenregel conservatief te kiezen. Deze conservatieve keuze was gebaseerd op de veronderstelling dat veel leerlingen eenvoudiger een voldoende konden halen voor wiskunde A dan voor wiskunde B. De keuze voor wiskunde C daalde in deze periode gestaag van 8% naar 3%.

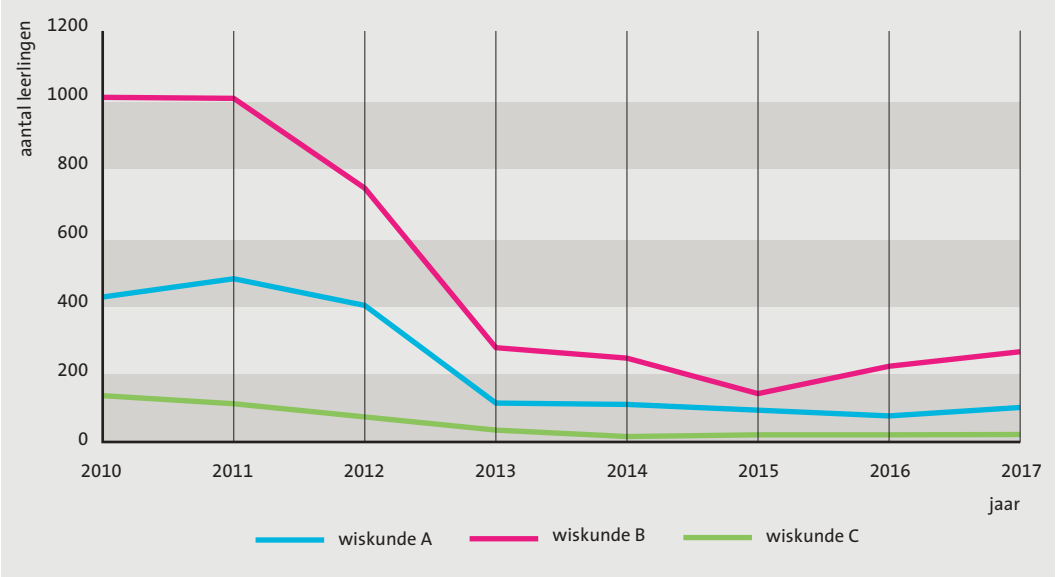


Figuur 6.15 Percentage vwo-examenkandidaten met wiskunde A en wiskunde B in het pakket (BRON, DUO)

6.2.3 Gevolgen van de kernvakkenregeling

Zakken

In 2013 werd de kernvakkenregel ingevoerd. Als gevolg daarvan moesten leerlingen voor wiskunde minimaal een 5 op hun eindlijst staan om een diploma te kunnen halen. Figuur 6.16 laat zien hoeveel leerlingen in de jaren 2010-2017 voor wiskunde een 4 of lager op hun eindlijst hadden.

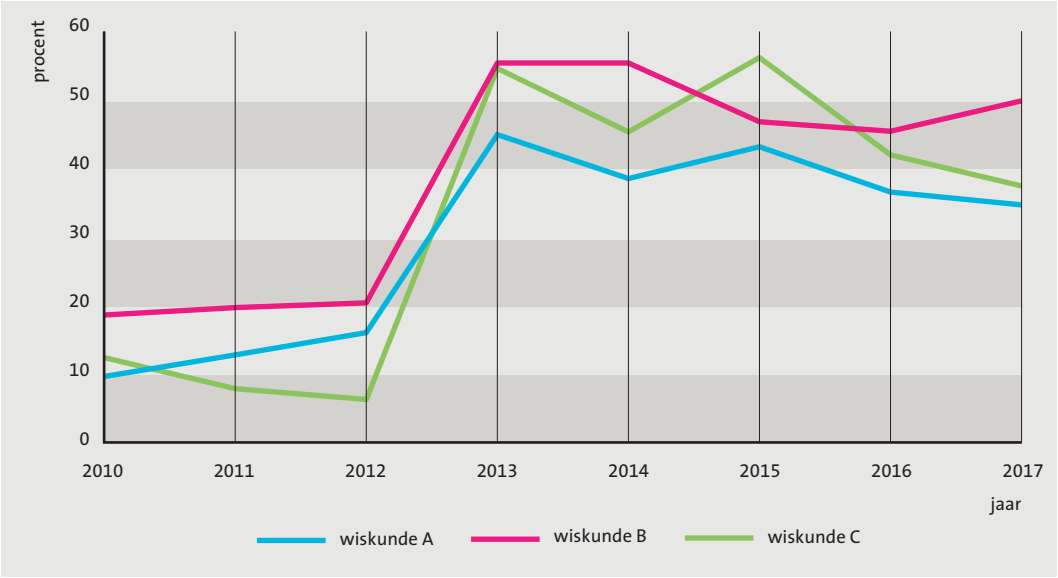


Figuur 6.16 Aantal vwo-leerlingen met een cijfer 4 of lager voor wiskunde op hun eindlijst (BRON, DUO)

Wat niet uit figuur 6.16 blijkt, maar wat we uit onderzoek wel weten, is dat vanaf 2013 het grootste deel van de leerlingen met een 4 voor wiskunde ook gezakt zou zijn zonder kernvakkenregeling. Het percentage leerlingen dat alleen zakte op basis van de 4 (of lager) voor wiskunde, daalde vanaf 2014 tot ver onder de 0,1%. Dat geeft aan dat scholen en leerlingen er goed in slagen te voorkomen dat er wordt gezakt op (alleen) een 4 of lager voor wiskunde.

Herkansen

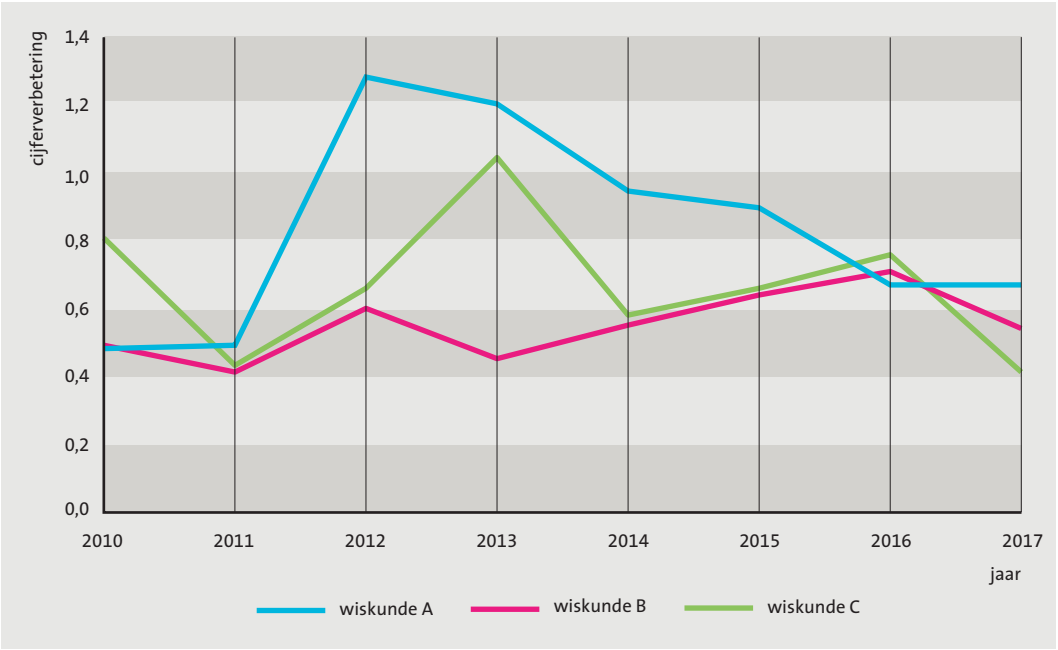
Figuur 6.16 geeft de situatie weer na afloop van het tweede tijdvak waarbij dus ook de herexamens zijn meegenomen. Maar hoe ziet de situatie eruit na het eerste tijdvak? En welke actie ondernemen leerlingen dan? In figuur 6.17 zijn alle leerlingen meegenomen met een gemiddeld cijfer van schoolexamen en eerste tijdvak centraal examen van lager dan 4,5. Zij zouden een 4 op hun eindlijst krijgen als ze niets zouden doen. Afgebeeld is het percentage leerlingen dat ervoor kiest om een herkansing wiskunde te doen. Opvallend is de sterke stijging sinds 2013. Waarschijnlijk kiezen leerlingen vanaf dat moment veel vaker voor een herkansing wiskunde, omdat dit de kans om alsnog te slagen groot maakt. Dit weerspiegelt het belang van wiskunde als kernvak.



Figuur 6.17 Percentage vwo-examenkandidaten met een 4 of lager voor wiskunde dat besluit om wiskunde te herkansen (BRON, DUO)

Ophalen

Het belang van een goed examencijfer voor wiskunde is na 2012 groter geworden. Dat zou betekenen dat vwo-leerlingen gemotiveerder zijn om de bijbehorende examens goed te maken. Interessant is of we dit terug zien komen in de gerealiseerde cijferverbetering in het tweede tijdvak. Scoren leerlingen na 2012 beter op hun herkansing dan daarvoor? Figuur 6.18 geeft aan hoeveel cijferpunten leerlingen zich gemiddeld wisten te verbeteren door het vak wiskunde te herkansen. De grafiek laat een grillig patroon zien, zonder duidelijke stijging vanaf 2013.

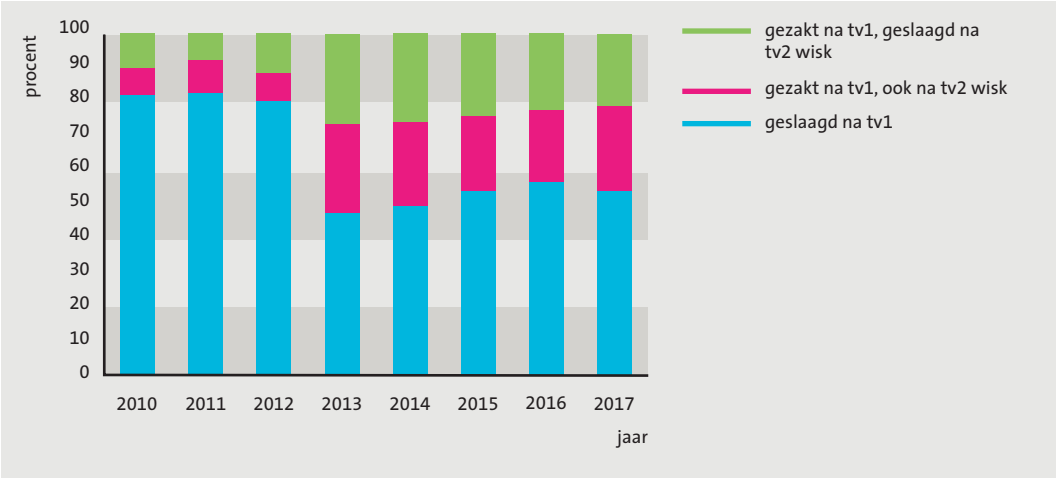


Figuur 6.18 Cijferverbetering vwo-examenkandidaten met een 4 of lager voor wiskunde na het eerste tijdvak, na de herkansing (BRON, DUO)

Dat de grafiek een grillig verloop laat zien, wordt veroorzaakt door de wijze waarop de N-term van het tweede tijdvak tot stand komt. In de meeste gevallen is de N-term van het tweede

tijdvak namelijk gelijk aan (of ligt deze dicht bij) de N-term van het eerste tijdvak. De moeilijkheid van het examen in het tweede tijdvak is echter niet altijd gelijk aan dat van het eerste tijdvak. Compensatie (ophoging van de N-term) vindt alleen plaats als de herkansing aanzienlijk moeilijker was dan het examen in het eerste tijdvak. Was het tweede tijdvak makkelijker, dan krijgen leerlingen – voor eenzelfde prestatie – een hoger cijfer. Op basis van de grafiek zouden we kunnen concluderen dat het tweede tijdvak voor wiskunde A in 2012, 2013, 2014 en 2015 veel gemakkelijker was dan het eerste tijdvak. Het tweede tijdvak voor wiskunde B was in 2011 en 2013 juist relatief moeilijk. Maar op basis van de grafiek zouden we net zo goed kunnen concluderen dat het voor wiskunde A-leerlingen eenvoudiger is om zich te verbeteren in het tweede tijdvak. Volgens deze redenering blijkt dat zij in de korte tijd tussen het eerste en tweede tijdvak meer bij kunnen leren. Nader onderzoek is nodig om dit leerwinst-effect en het effect als gevolg van verschil in moeilijkheid afzonderlijk te kunnen bepalen.

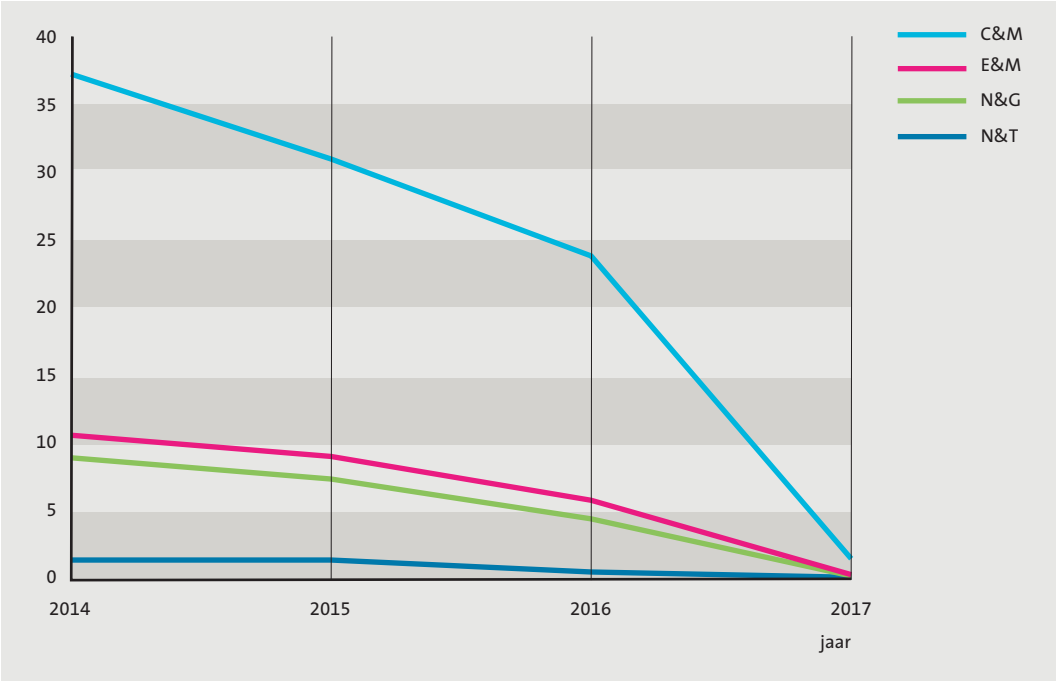
Mogen we veronderstellen dat het percentage examenkandidaten in de herkansing een 5 (of hoger) scoort, na 2013 toeneemt? Figuur 6.19 laat het zien. Als we alle wiskundevakken samen nemen, stijgt het aandeel leerlingen dat na een herkansing wiskunde alsnog slaagt van ongeveer 10% in de jaren 2010-2012 naar 20-25% in de jaren 2013-2017.



Figuur 6.19 Resultaat van de herkansing bij vwo examenkandidaten met een 4 of lager voor wiskunde na het eerste tijdvak (BRON, DUO)

6.3 De rekentoets

Hier volgt nog een overtuigend voorbeeld van de invloed van het belang van de toets. In 2017 telde de rekentoets voor het eerst (en laatst) mee in de kernvakkenregeling op het vwo. Doordat leerlingen slechts één vijf op hun eindlijst mochten hebben voor de kernvakken, waren leerlingen in 2017 zeer gemotiveerd om geen onvoldoende te scoren op de rekentoets. Afgaand op de resultaten in de jaren tot en met 2016 werd gevreesd dat vooral veel vwo-leerlingen met een C&M-profiel zouden zakken als gevolg van hun resultaat op de rekentoets. Figuur 6.20 geeft per profiel het percentage leerlingen weer dat een onvoldoende scoorde op de rekentoets. Wat blijkt is dat het percentage C&M-leerlingen met een onvoldoende voor de rekentoets van 2016 naar 2017 daalde van 24% naar 1%. Ook voor de andere profielen was de daling bijzonder groot (DUO, examenmonitor 2017).



Figuur 6.20 Percentage vwo-examenkandidaten met een onvoldoende op de rekentoets per profiel.

6.4 Samenvatting en conclusie

Toetsresultaten worden bepaald door cognitieve- en niet-cognitieve kenmerken. Eén van de niet-cognitieve factoren die een rol kan spelen, is motivatie. Zeker bij toetsen waar geen persoonlijke consequenties aan verbonden zijn, zijn leerlingen niet per se maximaal gemotiveerd een optimale toetsprestatie te leveren.

Low-stakes toetsen zoals PISA zijn gevoelig voor een lagere motivatie van leerlingen en dat kan effect hebben op de uiteindelijke resultaten. Dit hoofdstuk heeft geprobeerd vertekeningen door verschillen in motivatie voor beide vergelijkingen te adresseren voor Nederlandse PISA-resultaten. Omdat er geen directe meting van motivatie binnen PISA beschikbaar is, is er gebruik gemaakt van de prestatiemotivatie-index, responstijden en item nonresponse.

Nederland onderscheidt zich ten opzichte van andere landen vooral in de snelheid van antwoorden. Binnen de EU zijn Nederlandse leerlingen het snelst met het geven van antwoorden op de vragen in de PISA-toets. Dat geldt voor alle hoofdonderdelen van de toets. Nederlandse leerlingen onderscheiden zich ook ten opzichte van met name Oost-Aziatische leerlingen doordat de tijd die zij besteden aan opgaven een stuk meer afneemt naarmate de toets vordert. In combinatie met een afnemend aantal goed op dezelfde opgaven in de verschillende clusters in de toets in Nederland, terwijl deze relatie constant is in Oost-Azië, is een behoorlijk sterke indicatie dat leerlingen in Nederland minder moeite in de toets kunnen/willen investeren in vergelijking met leerlingen in Oost-Azië. Er lijkt dus een verschil in (intrinsieke) motivatie te bestaan tussen deze twee groepen leerlingen.

In vergelijking met de beoordeling van verschillen tussen landen, is er minder instrumentarium beschikbaar voor het evalueren van eventuele verschillen in motivatie tussen de verschillende afnames van PISA. De prestatiemotivatie-index is pas voor de eerste keer opgenomen in PISA-2015 en ook responstijden zijn pas voor de eerste keer in 2015 geregistreerd met de

introductie van een digitale afname van de toetsen. Na de afname van PISA-2018 kan dit wel uitgebreider onderzocht worden. Het aantal niet-ingevulde vragen en dan met name het aantal vragen dat aan het eind van de toets niet is ingevuld, is dan ook de enige indicator die gebruikt kan worden om verschillen in motivatie tussen PISA-2012 en PISA-2015 uit te drukken.
Het percentage niet- of verkeerd ingevulde vragen is niet wezenlijk veranderd in 2015 ten opzichte van 2012. Daarmee is er ook geen aanwijzing dat de trendresultaten binnen Nederland vertekend zijn door een verandering in motivatie.

Hoe motivatie een high-stake toets beïnvloedt, is bekeken aan de hand van het vwo-examen wiskunde. Na de introductie van de kernvakkenregeling is het aantal leerlingen dat een vier of lager op zijn eindlijst krijgt voor wiskunde gedaald. Er zakken minder leerlingen door de kernvakkenregel dan werd verwacht op basis van de resultaten in 2010 – 2012. Hiervoor zijn drie oorzaken aan te wijzen. Ten eerste zijn leerlingen sinds 2013 vaardiger geworden en halen zij hogere cijfers op hun schoolexamen (zie bijlage 1). Ten tweede halen zij hogere cijfers op hun centraal examen (zie paragraaf 2.2.3) En ten derde is het gedrag van leerlingen rond de herkansing een factor. Leerlingen herkansen vaker voor het vak wiskunde en het aantal leerlingen wat hierdoor alsnog zijn diploma haalt, is sinds 2013 gestegen. Ook voor de rekentoets geldt dat de eisen die gesteld worden in de zak/slaagregeling een regulerend effect hebben. Op het moment dat het resultaat zwaar meetelt, neemt het aantal onvoldoendes drastisch af. Scholen en leerlingen zijn kennelijk in staat hier nauwkeurig op in te spelen.
Al met al hebben scholen en leerlingen zich zodanig aan de nieuwe eisen en regels aangepast dat de verwachting dat een groot aantal kandidaten zou zakken op basis van de kernvakken-regel niet is uitgekomen.

7 Slotbeschouwing

7 Slotbeschouwing

In dit rapport zijn de onderzoeksvragen zoals benoemd in hoofdstuk 1 nader uitgewerkt. In dit hoofdstuk vatten we de resultaten samen en geven we enkele beperkingen aan. Ook gaan we in op de context van dit onderzoek, namelijk op aspecten van de praktijk van eindexamens. De commissie Steur heeft Cito gevraagd om een inhoudelijke reactie te geven vanuit haar rol expertisecentrum op het gebied van examinering en van psychometrisch onderzoek. Hiermee besluiten we dit hoofdstuk.

7.1 De onderzoeksresultaten per hoofdstuk

De eerste onderzoeksvraag van hoofdstuk 2 betreft de vaardigheidsontwikkeling van in de PISA-onderzoeken en de factoren die hierop van invloed waren. De prestaties van Nederlandse vo-scholieren op het internationale PISA-onderzoek laten een divers beeld zien. Voor leesvaardigheid blijven de prestaties op een constant niveau. Voor wiskundige geletterdheid zien we een vrij constante, lichte daling. Tussen 2012 en 2015 speelt deze overigens alleen bij havo/vwo en vmbo-bb. De prestaties bij natuurwetenschappelijke geletterdheid, ten slotte, blijven tot en met 2012 constant en laten dan in 2015 een abrupte daling zien. Die daling geldt niet alleen voor Nederland, maar gemiddeld ook voor andere OESO-landen. Met name vmbo-leerlingen deden het dat jaar minder goed. Misschien omdat de nieuwe definitie (met de nadruk op wetenschappelijk onderzoek) minder aansluit bij hetgeen 15-jarige vmbo-leerlingen leren op school. Maar misschien ook omdat de digitale afname of nieuwe, interactieve, opgaven minder goed aansluiten bij hun digitale vaardigheden. Het uitvoerend consortium van PISA heeft geprobeerd in de analyse rekening te houden met deze wijzingen, zodat ze geen invloed hebben op de trendvergelijkingen. Desondanks moeten de trendresultaten met meer dan de gebruikelijke voorzichtigheid geïnterpreteerd worden.

De tweede onderzoeksvraag van hoofdstuk 2 betreft de cijferontwikkeling op centraal examen en schoolexamen van de relevante vakken en de factoren die van invloed waren. Bij de centrale examens zijn er behoorlijk wat wijzigingen in de examenregeling en het examenprogramma geweest. Al deze wijzigingen zien we amper terug in de cijfers van het schoolexamen. Wel zien we dat 2011 een keerpunt is. De wijzigingen zijn daarentegen wel zichtbaar bij de centrale examens, waar 2011 ook een keerpunt was. Het gemiddelde eindexamencijfer kent tot 2011 een dalende trend. Vanaf 2012 (aanscherping exameneis) blijft de trend in het vmbo overwegend stijgend. In havo/vwo stijgt het cijfer in 2012 (aanscherping exameneis) en 2013 (introductie kernvakkenregeling). In 2014 daalt het gemiddelde cijfer om daarna ongeveer gelijk te blijven. Specifiek bij Nederlands havo/vwo zien we een daling in 2015, waar de ijking aan de referentieniveaus leidde tot een verzwaring van de prestatie-eis bij vwo. Bij wiskunde slaat de stijging die na 2011 is ingezet, in vmbo-kb, vmbo-gt/tl en havo na 2015 om in een duidelijke daling. Engels wijkt af van het algemene beeld doordat de trend op alle niveaus gedurende de hele periode 2012-2017 overwegend stijgend blijft.

Hoofdstuk 3 beschrijft hoe de N-term tot stand komt en welke rol normhandhavingsonderzoek daarin speelt. Normhandhavingsonderzoeken spelen een belangrijke rol bij de berekening van de technische N-termen. De definitieve N-term wijkt in een beperkt aantal gevallen af van de technische N-term. Bij een op de zes vakken wordt de N-term met gemiddeld 0,1 verhoogd. In de meeste gevallen is deze ophoging omdat er inhoudelijk iets aan de hand was met een vraag. De gemiddelde cijfers van de centrale examens komen bij wiskunde, Engels, biologie, natuurkunde en economie grotendeels overeen met de vaardigheidsontwikkeling die uit de gegevens van de normhandhavingsonderzoeken spreekt. Enkele onverwachte vaardigheids-stijgingen zijn in de praktijk niet gevolgd. Dat heeft te maken met de demping, die is ingesteld omdat resultaten van normeringsonderzoek door toevalsfluctuaties soms te extreem kunnen uitvallen om geloofwaardig te zijn. Bij een grote toe- of afname van de vaardigheid ten opzichte van de referentiepopulatie wordt de N-term conservatief vastgesteld.

Hoofdstuk 4 gaat in op de inhoudelijke verschillen tussen PISA en examenprogramma's, en de operationalisatie in de opgaven. De focus en interpretatie van 'leesvaardigheid' verschilt in sommige opzichten fundamenteel. Terwijl sommige PISA-vragen in het domein leesvaardigheid zouden kunnen voorkomen in de vmbo-examens Nederlands, kan een groot deel van de examenvragen niet voorkomen in PISA, met name vanwege het tekstbeschouwelijke karakter. De examenvragen Nederlands zijn ook abstracter en complexer. Wiskundige geletterdheid zoals dat voorkomt in PISA, is voorbereidend op het examenprogramma. Het is echter zo basaal dat het moeilijk te herkennen is in de diverse examenprogramma's. De vraagvorm en scoringswijze verschillen ook tussen PISA en de eindexamens.
Tussen natuurwetenschappelijke geletterdheid in PISA en getoetste vakkennis bij de eind-examens is een verschil. De keuze van de contexten is geheel anders: waar PISA-contexten in het teken staan van burgerschap, staan de contexten in de examens in het teken van vakmanschap.
Bij alle drie onderdelen zijn de verschillen tussen PISA-inhouden en exameninhouden zo groot dat een zinvolle vergelijking niet goed mogelijk is.

In hoofdstuk 5 gaan we in op de wijzigingen in leerlingenstromen. De doorstroom van leerlingen over de jaren 2008-2017 is niet constant. De doorstroom fluctueert licht, met een zichtbaar keerpunt in 2011. Het keerpunt 2011 in de stijgende trend van zittenblijven en afstroom, en de graduele stijging in het niet-aanmelden voor het examen vanaf 2014, zijn opvallend. In hoofdstuk 2 zien we ook bij de gemiddelde cijfers een keerpunt na 2011. Met de aanscherping van de exameneis in 2012 begint het gemiddelde cijfer op het centraal eindexamen te stijgen. Vanaf 2012 is een daling ingezet in zittenblijven en de afstroom, wat een verdere stijging van de examencijfers in de daarop volgende jaren kan hebben afgezwakt.

In hoofdstuk 6 staat de vraag centraal welke invloed het belang van toetsing en motivatie hebben op prestaties. Low-stakes toetsen zoals PISA zijn gevoelig voor een lagere motivatie van leerlingen en dat kan effect hebben op de uiteindelijke resultaten. Er is echter geen aanwijzing dat verschillen in motivatie trendvergelijkingen vertekenen. Er zijn wel redelijk sterke indicaties dat leerlingen in Nederland minder bereid zijn moeite te stoppen in het oplossen van opgaven dan Oost-Aziatische leerlingen naarmate de toets vordert.
Bij de centrale eindexamens zien we dat het belang van centrale eindexamens, met name van de kernvakken in havo/vwo, is toegenomen in de diplomabeslissing. Scholen en leerlingen hebben zich aangepast aan de nieuwe eisen en regels. Gemiddeld genomen presteren leerlingen beter op het centrale examen wiskunde-vwo na de invoering van de kernvakken-regeling. Doordat leerlingen vaker voor een herkansing wiskunde kiezen, neemt het gemiddelde cijfer nog verder toe. Er zakken minder leerlingen door de kernvakkenregel dan werd verwacht op basis van de resultaten in 2010 – 2012.

7.2 Beperkingen van het onderzoek

Het uitgevoerde onderzoek kent beperkingen. Zo wordt een directe vergelijking van prestaties op eindexamens en PISA-toetsen verhinderd doordat de PISA-kandidaten geanonimiseerd zijn. Een koppeling van de scores van geanonimiseerde PISA-leerlingen aan hun eigen eindexamen-prestaties in de daarop volgende jaren, is onmogelijk. We kunnen dus alleen indirect de trends in vaardigheidsontwikkeling in PISA en eindexamens beschrijven, vergelijken en verklaren.

In dit onderzoek wordt ook geen aandacht besteed aan de verschillen tussen jongens en meisjes. Het gaat in dit onderzoek om factoren die het zicht op de vaardigheidsontwikkeling bevorderen of bemoeilijken. Daardoor is de norm voor bevordering naar de examenklas wel een factor, omdat deze mede bepaalt hoe vaardig de groep leerlingen gemiddeld is. De verdeling van de examenpopulatie in jongens en meisjes is op zich interessant, maar vormt geen bepalende factor voor het niveau van de groep.

Uit de gegevens van DUO kan de precieze schoolcarrière van individuele leerlingen niet worden afgeleid. De relatie tussen schoolcarrière en examencijfer kunnen we dus niet direct vaststellen. We beschrijven alleen de trend in zittenblijven en afstroom in de bovenbouw als geheel. In dit onderzoek zijn ook trends in het aantal leerlingen dat het vo tussentijds verlaat niet nader onderzocht, omdat niet duidelijk is hoeveel leerlingen dit betreft.
Cito heeft geen nadere gegevens omtrent schoolcarrières opgevraagd bij andere partijen, zoals bijvoorbeeld de Inspectie van het Onderwijs. Alleen openbaar gepubliceerde en reeds aan ons verstrekte gegevens zijn benut. Relevante andere informatie zou de relatie tussen schoolcarrière en eindexamencijfers wellicht volledi ger in kaart kunnen brengen.

Bij PISA ontbreekt een directe meting van motivatie op meerdere meetmomenten. Noch de relatie tussen motivatie en prestatie op individu-niveau, noch trends in motivatie bij PISA hebben we dus onderzocht.

7.3 Robuustheid van het systeem van examinering

De motie spreekt over de rol van N-termen bij de normering van centrale examens. Centrale examens vormen een deel van het gehele systeem van examinering. Om iets over de robuustheid van het systeem te kunnen zeggen, is het nodig om iets te zeggen over de nauwkeurigheid van de cijfers, die op hun beurt iets kunnen zeggen over de nauwkeurigheid van de diplomabeslissing.

In de aanloop naar de diplomabeslissing worden de cijfers van schoolexamens en centrale examens tot op een decimaal nauwkeurig weergegeven. Twee factoren bepalen de nauwkeurigheid van het cijfer: de meetonzekerheid (=betrouwbaarheid) van het examen en de nauwkeurigheid van de normoverbrenging.

Elk examen bestaat uit een strak gereguleerde set van vragen. Hypothetisch gezien is deze set een greep uit alle mogelijke sets die ook het examen hadden kunnen vormen. Afhankelijk van de specifieke vragen die in het examen worden opgenomen, kan het examen iets makkelijker of iets moeilijker uitvallen. Voor deze fluctuaties in moeilijkheid wordt gecorrigeerd in de normering. Maar voor een individuele leerling kunnen sommige vragen net wat makkelijker of moeilijker zijn doordat sommige vragen net iets beter aansluiten bij de onderdelen die de leerling goed beheerst. Dit is een van de belangrijke oorzaken van de meetonzekerheid bij een examen. De ordegrootte van die meetonzekerheid ligt meestal rond de 0,5 cijferpunt. Als we deze meetonzekerheid willen verkleinen, moeten we het examen substantieel langer maken.

De nauwkeurigheid van het cijfer hangt daarnaast samen met de nauwkeurigheid van de normoverbrenging. Deze nauwkeurigheid is het grootst, wanneer een deel van dat examen

exact gelijk is aan een deel uit een vroeger examen. Natuurlijk moet dit deel van het examen dan geheim zijn gebleven. Met deze vorm van normeren (die we toepassen bij de digitale centrale examens in vmbo bb en vmbo kb) kunnen we de norm op één decimaal nauwkeurig overbrengen. Iets minder nauwkeurig is de methode van post- en pretesten. Toch levert deze normoverbrenging relevante informatie op, omdat de onnauwkeurigheid van deze methoden nog steeds veel kleiner is dan de meetonzekerheid van het examen zelf.

Deze overwegingen leiden ertoe dat het cijfer achter de komma van de examencijfers met de nodige voorzichtigheid bekeken moet worden. In het systeem van examinering wordt hieraan tegemoet gekomen: er is uitmiddeling tussen school- en centraal examen, en leerlingen krijgen de mogelijkheid tot herkansing. Ten slotte geldt de regel dat leerlingen met beperkte onvoldoendes toch een diploma kunnen krijgen. Dit vergroot de kans dat een leerling die door de meetonzekerheid een onvoldoende kreeg, maar die wel over voldoende vaardigheid beschikt, toch zijn diploma krijgt. Zo zijn er voldoende verzekeringen in het systeem ingebouwd om te zorgen dat een diplomawaardige leerling het diploma ook echt krijgt.

De onzekerheid van de normering van de centrale examens is in dit rapport uitgebreid belicht. Het is goed om dit in de bredere context van de examinering te plaatsen. Zolang we in Nederland waarde hechten aan vakkennis en aan borging van het niveau, blijven de centrale examens een belangrijke mijlpaal in de onderwijscarrière van leerlingen.

7.4 Verbetering van de normhandhaving

Dit rapport behandelt uitgebreid de normering van centrale examens en de nauwkeurigheid daarvan. De wijze van normhandhaving is geen statisch proces. De wereld verandert en de normering moet ook mee veranderen. Cito zoekt daarbij continu naar mogelijkheden tot verbetering.

Voortschrijdende techniek of een nieuwe theorie bieden soms nieuwe kansen om de normhandhaving te verbeteren. Ook verandert de maatschappij, het onderwijssysteem of de wensen omtrent normering, wat aanpassingen van de methodiek vereist. De introductie van de Fishermethode was bijvoorbeeld een gevolg van de invoering van de verzwaarde exameneisen. Tot slot is er altijd voortschrijdend inzicht. Normeren blijft mensenwerk en de ervaringen met de werkwijzen waarmee genormeerd wordt, zijn voortdurend onderwerp van evaluatie en reflectie. De vereiste zorgvuldigheid maakt dit noodzakelijk.

In de loop der jaren zijn er verbeteringen en aanvullende methoden geweest. Dat neemt niet weg, dat het wijzigen van de wijze van normhandhaving gevoelig ligt. Onterechte twijfels over de juistheid van de gehanteerde wijze in de voorgaande jaren kunnen dan opborrelen, of vernieuwingen worden als overbodig beschouwd. Normhandhaving is daarom van nature conservatief en daarbij gaan we niet te lichtvoetig tot wijziging over. Toch, in retrospectief, en kijkend naar een langere periode, kregen leerlingen in de jaren ‘80 hun diploma op basis van een totaal andere wijze van normhandhaving dan vandaag. Maar dat maakt hun diploma niet minder waard.

Ook op dit moment denken we na over verbetering van de normering. Daarbij zijn een aantal thema’s in onderzoek. Een eerste thema is vereenvoudiging. De zeer korte tijd waarin er jaarlijks genormeerd moet worden, vraagt om een eenvoudige en doeltreffende werkwijze. Eenvoudige werkwijzen hebben een lager risico op procedurele fouten. Ook kunnen vereenvoudigde werkwijzen minder tijd vereisen, waardoor bijvoorbeeld de deadline voor docenten voor het inleveren van de scores in Wolf naar achteren kan worden verplaatst. Docenten krijgen zo meer tijd om te corrigeren, wat hun werkdruk verlicht en de beoordeling van de leerlingen ten goede komt.

Een tweede thema is verbreding van de referentiepopulatie waarmee prestaties van nieuwe leerlingen worden vergeleken. In zijn brief van januari 2019 (Kamerbrief over toetsing en examinering in het voortgezet onderwijs 2018, januari 2019) maakt de minister duidelijk dat met ingang van 2019 gekozen wordt voor een bredere basis onder de referentiegegevens. De nieuwe normeringsmethodiek vooronderstelt dat de referentiegegevens stabieler zijn wanneer meerdere jaren worden meegenomen in de bepaling van deze referentiegegevens. Een derde thema is het tweede tijdvak. De minister heeft in dezelfde brief gevraagd om een verkenning van een eigenstandige normering voor het tweede tijdvak en de gevolgen daarvan.

Een vierde thema is transparantie. Het normeringsproces is soms complex en lijkt dan ondoorzichtig, hetgeen niet direct bijdraagt tot vertrouwen in het proces. We zullen als Cito en CvTE duidelijker moeten maken hoe de N-term tot stand is gekomen met aandacht voor de berekening van de technische N-term door Cito en de vaststelling van de definitieve N-term door CvTE met eventuele verwerking van bijzondere omstandigheden. Het publiceren van de resultaten van normhandhavingsonderzoeken is daarbij een punt van zorg. Betrokken partijen wensen soms meer nauwkeurigheid van het vaststellingsproces van de N-term dan praktisch haalbaar is met de beschikbare middelen. Het alternatief is om de resultaten van norm-handhavingsonderzoeken te plaatsen in een trend van meerdere jaren. Op deze manier is de uitkomst minder gevoelig voor toevallige schommelingen van de techniek.

Een laatste thema zijn de verschillen tussen vakken. Het gemiddelde cijfer tussen vakken, verschilt soms aanzienlijk. Dat heeft te maken met twee zaken. Zo werd rond de eeuwwisseling voor elk vak afzondelijk een gewenste norm vastgesteld. Daar hoorde dan een gemiddeld cijfer en een percentage onvoldoende bij. Deze norm is in de loop van de jaren gehandhaafd. Daarnaast verschuift het gemiddelde cijfer bij vakken waar een vaardigheidsontwikkeling plaatsvindt.
De gemiddelde cijfers voor bijvoorbeeld Engels en wiskunde B in VWO lagen acht jaar geleden rond de 6,5. De laatste jaren is dat een 7,0. De verklaring ligt (deels) in de invoering van de kern-vakkenregeling. De eeuwwisseling is inmiddels bijna 20 jaar geleden. Sommige vakverenigingen pleiten voor een herbezinning op de huidige normen (Kamerbrief over toetsing en examinering in het voortgezet onderwijs 2018, januari 2019). In deze Kamerbrief noemt de Minister havo Frans en Duits. Hij vraagt het CvTE een analyse te maken en de wenselijkheid te onderzoeken om de eisen aan de veranderde vaardigheid aan te passen. Ook vraagt hij wat de mogelijke oorzaak kan zijn van die veranderde vaardigheid.

7.5 Tot slot

De trends in PISA en centrale examens zijn verschillend. In dit rapport hebben we veel nuanceringen aangebracht en hebben we regelmatig opgeroepen om voorzichtig te zijn met het trekken van conclusies. Er spelen veel factoren een rol bij het verklaren van de verschillen in de trends. Dit maakt de analyse complex. Het is niet zo dat er één factor is die het trendverschil kan verklaren. Maar als we een factor mogen noemen die in onze ogen een belangrijke rol speelt, dan is dat het toegenomen belang van de centrale examens voor zowel scholen als voor leerlingen. Scholen hechten aan goede examenresultaten, sturen de leerlingenstromen en besteden veel aandacht aan de centrale examens. Het rapport van de commissie kwaliteit schoolexamens (Ten Dam, 2018) laat dit duidelijk zien. Voor leerlingen vertaalt het grotere gewicht van de centrale examens in de zak/slaagregeling sinds 2013, zich in meer aandacht voor de examenvoorbereiding. Denk daarbij bijvoorbeeld aan het volgen van georganiseerde examentrainingen (De Geus & Bisschop, 2018). Dit alles mist zijn uitwerking niet: de prestaties op de eindexamens zijn verbeterd.

Literatuurlijst

Literatuurlijst

Adams, R. & Wu, M. (eds.) (2002). *PISA 2000 technical report*. Parijs: OECD.

Butler, J., & Adams, R. (2007). *The impact of student effort on the outcomes of international studies*. Journal of Applied Measurement, 8 (3), 279-304.

Cito (2004). *Resultaten PISA-2003: Praktische kennis en vaardigheden van 15 jarigen*. Arnhem: Cito.

Cito (2007). *Resultaten PISA-2006: Praktische kennis en vaardigheden van 15-jarigen*. Arnhem: Cito.

Cito (2010). *Resultaten PISA-2009: Praktische kennis en vaardigheden van 15-jarigen*. Arnhem: Cito.

Cito (2013). *Resultaten PISA-2012: Praktische kennis en vaardigheden van 15-jarigen*. Arnhem: Cito.

Cito (2016). *Resultaten PISA-2015: Praktische kennis en vaardigheden van 15-jarigen*. Arnhem: Cito.

Davidov, E. Schmidt, P, Billiet, J. & Meuleman, B. (Eds.), (2018) *Cross-cultural analysis: Methods and applications,* 2nd ed. London: Taylor and Francis.

De Geus, W. & Bisschop, P. (2018). *Licht op schaduwonderwijs. Onderzoek naar deelname aan en uitgaven voor schaduwonderwijs.* SEO & Oberon. URL: http://www.seo.nl/uploads/media/2018-09_Licht_op_schaduwonderwijs.pdf

DUO (2012). *Examenmonitor VO 2012*. Groningen: DUO.

DUO (2013). *Examenmonitor VO 2013*. Groningen: DUO.

DUO (2014). *Examenmonitor VO 2014*. Groningen: DUO.

DUO (2015). *Examenmonitor VO 2015*. Groningen: DUO.

DUO (2016). *Examenmonitor VO 2016*. Groningen: DUO.

DUO (2017). *Examenmonitor VO 2017*. Groningen: DUO.

Feskens, R., & Koops, J. (2016). *PISA 2015: Evaluation of nonresponse in the Netherlands (ongepubliceerd manuscript).* Arnhem: Cito.

Feskens, R.C.W., Fox, J.P. & Zwitser, R. (in press). *Differential Item Functioning in PISA due to mode effects.*

Gneezy, U., List, J.A., Livingston, J.A., Sadoff, S., Qin, X., Xu, Y.(2017). *Measuring success in education: the role of effort on the test itself*. National Bureau of Economic Research: Working Paper 24004.URL: http://www.nber.org/papers/w24004

Holland, P.W., & Wightman, L.E. (1982). *Section pre-equating: A preliminary investigation.* In P.W. Holland & D.R. Rubin (Eds.), Test Equating (pp. 271-297). New York: Academic Press.

Hopfenbeck, T.N. & Kjærnsli. M. (2016). *Students' test motivation in PISA: the case of Norway.* The Curriculum Journal, 27 (3), 406-422, DOI: 10.1080/09585176.2016.1156004

Keizer-Mittelhaëuser, M. (2014). *Modeling the effect of differential motivation on linking educational tests.* Tilburg University: PhD thesis.

Kiplinger, V.L., & Linn, R. L. (1996). *Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress.* Educational Assessment, 3, 111-133.

Kolen, M.J. & Brennan, R.L. (2014). *Test equating, scaling, and linking.* Springer New York, 2014. DOI: 10.1007/978-1-4939-0317-7.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). *Estimating population characteristics from sparse matrix samples of item responses.* Journal of Educational Measurement, 29, 133-161.

O'Neill, H.F., Sugrue, B., & Baker, E.L. (1996). *Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance.* Educational Assessment, 3, 135-157.

OCW, DUO, & CBS. (2019). *Onderwijs in cijfers.* Geraadpleegd van https://www.onderwijsincijfers.nl/kengetallen/vo/leerlingen-vo

OECD (2000). Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy. Parijs: OECD.

OECD (2001). *Knowledge and skills for life. First results from PISA 2000.* Paris: OECD.

OECD (2002). *PISA 2000 technical report.* Parijs: OECD.

OECD (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills.* Parijs: OECD.

OECD (2005). *PISA 2003 technical report.* Parijs: OECD.

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006.* Parijs: OECD.

OECD (2009). *PISA 2006 technical report.* Parijs: OECD.

OECD (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science.* Parijs: OECD.

OECD (2012). *PISA 2009 technical report.* Parijs: OECD.

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* Parijs: OECD. http://dx.doi.org/10.1787/9789264190511-en

OECD (2014). *PISA 2012 technical report.* Paris: OECD.
http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education.* Parijs: OECD.
http://dx.doi.org/10.1787/9789264266490-en.

OECD (2017a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving (revised edition).* Parijs: OECD.
http://dx.doi.org/10.1787/9789264281820-en.

OECD (2017b). *PISA 2015 technical report.* Paris: OECD.
https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf.

R Core Team. (2016). *R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.* http://www.R-project.org/.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.H. (2017). *Herausforderungen bei der schätzung von Trends in Schulleistungsstudien.* Diagnostica, 63(2), 148–165. DOI: 10.1026/0012-1924/a000177.

Rubin, D. B. (Ed.) (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Ten Dam, G., Béguin, A., Waslander, S. & Van Leeuwen, M. (2018). *Een volwaardig schoolexamen. Rapport van de Commissie Kwaliteit Schoolexaminering.* Commissie Kwaliteit Schoolexaminering.
https://www.vo-raad.nl/system/downloads/attachments/000/000/696/original/Een_volwaardig_schoolexamen_Rapport_Commissie_Kwaliteit_Schoolexaminering_17_december_2019.pdf?1545036000

Von Davier, M. Gonzalez, E. & Mislevy, R.J. (2009). *What are plausible values and why are they useful? IERI monograph series: issues and methodologies in large-scale assessments.* Hamburg: IERI.

Wise, S. L., & Kong, X. (2005). *Response time effort: A new measure of examinee motivation in computer-based tests.* Applied Measurement in Education, 18, 163-183.

Wise, S.L. & DeMars, C.E. (2005). *Low examinee effort in low-stakes assessment: Problems and potential solutions.* Educational Assessment, 10, 1-17.

Wise, S.L. (2007). *Examinee effort and test score validity. Paper presented at the annual meeting of the Northeastern Educational Research Association.* Rocky Hill, Connecticut.

Wolf, L. F., & Smith, J. K., & Birnbaum, M.E. (1995). *The consequence of performance, test, motivation, and mentally taxing.* Applied Measurement in Education, 8, 341-351.

Wu, M. (2005). *The role of plausible values in large-scale surveys.* Studies in Educational Evaluation, 31, 114-128.

Zamarro, G., Hitt, C., and Mendez, I. (2016). *When students don't care: Reexamining international differences in achievement and non-cognitive skills.* EDRE Working Paper No. 2016-18. Arkansas: EDRE. DOI: 10.2139/ssrn.2857243

Begrippenlijst

Begrippenlijst

1e tijdvak	Eerste afnameperiode centrale schriftelijke examens in mei.
2e tijdvak	Tweede afnameperiode centrale schriftelijke examens in juni (zie ook herkansing en herexamen).
afstroom	Leerlingen die na een jaarovergang het onderwijs vervolgen in een lagere leerweg of een lager schooltype.
AIP	Anchor in Package. Normhandhavingstechniek waarbij ankeropgaven hergebruikt worden in meerdere examens.
Anchor in Package	Zie AIP.
ankeritem	Opgave die in twee of meer toetsen is opgenomen en gebruikt wordt om scores te equivaleren.
betrouwbaarheid	Mate waarin scores consistent, nauwkeurig en reproduceerbaar zijn, dat wil zeggen vrij van meetonzekerheden.
betrouwbaarheidsinterval	Interval dat met een zekere kans de populatieparameter omvat.
cba	Computer based assessment.
cbt	Computer based testing.
CE	Centraal examen.
CE-eis	Het gemiddelde van de centrale examencijfers moet minstens 5,50 zijn (zak/slaagregeling 2012).
cesuur	Grens tussen de hoogste toetsscore waaraan een onvoldoende, en de laagste toetsscore waaraan een voldoende wordt toegekend.
cohort	Groep personen uit een bepaalde leeftijdsgroep.
cse	Centraal schriftelijk examen.
cspe	Centraal schriftelijk en praktisch examen.
cTwo	Commissie toekomst wiskunde onderwijs (nieuw programma havo/vwo m.i.v. 2015).
CvTE	College voor Toetsen en Examens.
DUO	Dienst Uitvoering Onderwijs.
equivaleren	Statistische procedure om scores van verschillende toetsen vergelijkbaar te maken door ze op dezelfde schaal te brengen.
examenlijn	Loket waar docenten vragen kunnen stellen en klachten kunnen indienen over de centrale examens.
Examenmonitor	Jaarlijks door DUO uitgegeven rapport dat een beeld geeft van de prestaties van de leerlingen op het eindexamen.
Fisher-methode	Meta-analysetechniek die wordt gebruikt bij het berekenen van de generieke vaardigheidsstijging. Hierbij wordt voor een groep van coherente vakken de resultaten van meerdere normhandhavingsonderzoeken gemiddeld, met een weging op basis van de nauwkeurigheid.
geletterdheid	Mate waarin de kandidaat in staat is de kennis en vaardigheden toe te passen in praktische en realistische contexten.
gesloten vragen	Vraagtype waarbij de kandidaat moet kiezen uit een beperkt aantal antwoordmogelijkheden die vooraf gegeven zijn.
goniometrie	Onderdeel van de wiskunde.
herexamen	Zie herkansing.

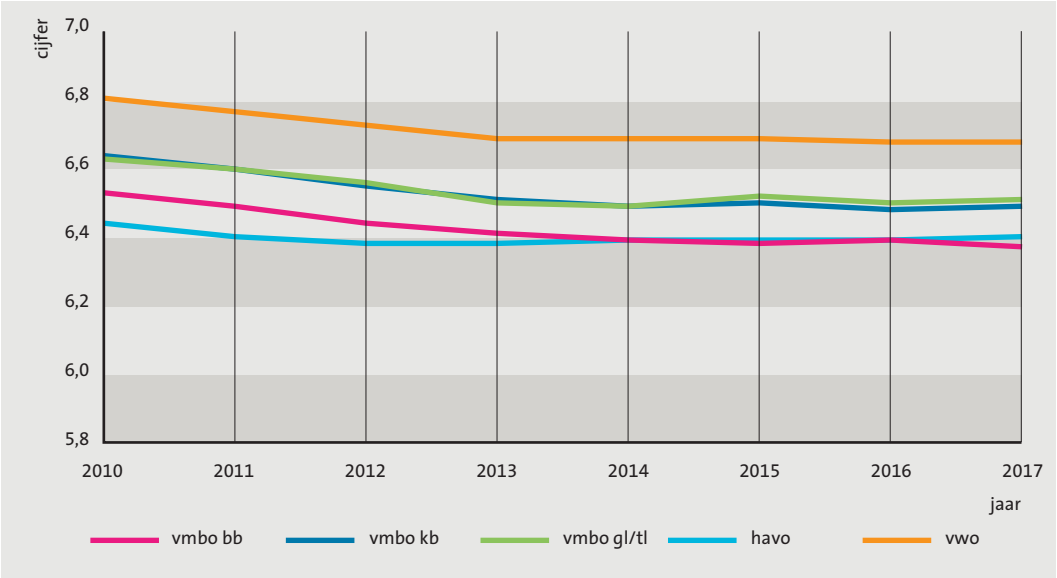
herkansing	Toegestane tweede zitting van een kandidaat, waarmee het cijfer van de eerste zitting verbeterd mag worden. Zie ook 2e tijdvak.
high-stakes toets	Toets waarbij er voor de kandidaat veel afhangt van het resultaat.
IRT	Itemresponstheorie. Statistische theorie binnen de psychometrie.
item	Synoniem voor opgave (vooral gebruikt bij meerkeuzetoetsen).
item parameter	IRT-kenmerk van een opgave, bijvoorbeeld de moeilijkheid.
kernvak	Vakken van belang voor de kernvakkenregel. Voor havo/vwo zijn Nederlands, Engels en wiskunde kernvakken, op vmbo is alleen Nederlands een kernvak. Op vwo telde in 2017 de rekentoets mee als kernvak.
kernvakkenregel	Regel binnen de zak/slaagregeling. Voor havo/vwo geldt dat een kandidaat minimaal een vijf moet hebben voor de kernvakken (2013). Voor vmbo geldt dat een kandidaat minimaal een vijf moet hebben voor Nederlands (2014).
LAKS	Landelijk Aktie Komitee Scholieren.
leerweg	Stroming binnen vmbo. Er zijn er vier: Basisberoepsgericht (bb), kaderberoepsgericht (kb), gemengd (gl) en theoretisch (tl).
Likert schaal	Wordt gebruikt om houdingen of opinies te meten. Personen worden verzocht te reageren op meerdere verwante items/ uitspraken/ stellingen en kiezen daarvoor een voorgegeven antwoordmogelijkheid (bijvoorbeeld: zeer mee eens... zeer mee oneens). Die elementen heten Likert-items.
low-stakes toets	Toets waarbij er voor de kandidaat weinig afhangt van het resultaat.
M&O	Management en organisatie. Vak op havo en vwo.
mediaan	Middelste waarde wanneer waarnemingen naar grootte gerangschikt zijn.
meerkeuze-item	Synoniem voor meerkeuzevraag. Vraag waarbij de kandidaat het goede antwoord moet bepalen uit verschillende alternatieven.
meetonzekerheid	Spreiding van het verschil tussen de geschatte vaardigheid en de werkelijke vaardigheid.
nask1	Natuur en scheikunde 1. Vak binnen vmbo waarin naast algemeen natuurwetenschappelijke onderwerpen vooral ook natuurkundige onderwerpen aan de orde komen.
normeren	Het bepalen van regels waarmee toetsscores in cijfers kunnen worden omgezet. Een belangrijk onderdeel van de normering is het vaststellen van de cesuur en daarmee de N-term.
normhandhaving	Procedure om scores, behaald bij verschillende examens, op dezelfde wijze te waarderen. Vaak via ankeritems of referentietoetsen.
normhandhavingsonderzoek	Onderzoek voor de normhandhaving.
NRO	Nationaal Regieorgaan onderwijsonderzoek.
N-term	Getal tussen 0 en 2 dat bij de omzetting van scores naar cijfers wordt gebruikt om de moeilijkheid van het examen te compenseren.
OCW	Ministerie van Onderwijs, Cultuur en Wetenschappen

OECD	Organisation for Economic Co-operation and Development. Zie ook OESO (Nederlandse afkorting).
OESO	Organisatie voor Economische Samenwerking en Ontwikkeling. Zie ook OECD (Engelse afkorting).
open vraag	Vraagtype waarbij een kandidaat het antwoord zelf moet formuleren.
opstroom	Leerlingen die na een jaarovergang het onderwijs vervolgen in een hogere leerweg of een hoger schooltype.
peilingsonderzoek	Grootschalig evaluatie-onderzoek om de inhoud en het rendement van (het onderwijs van) een bepaald schooltype vast te stellen.
pep	Platform examenprogramma's.
PISA	Programme for International Student Assessment.
plausible values	Set van aan aannemelijke waarden om de vaardigheid van leerlingen zuiver te schatten Wordt gebruikt in internationale onderwijskundige surveys, zoals PISA. Plausible values zijn geen traditionele individuele scores en kunnen daardoor ook niet gebruikt worden voor individuele rapportages.
posttest	Normhandhavingstechniek bij centrale examens. Na afloop van de examenafname wordt een aantal examenopgaven samen met ankervragen voorgelegd aan een groep personen.
PPON	Periodieke Peiling van het Onderwijsniveau.
pretest	Normhandhavingstechniek bij centrale examens. Ruim voorafgaand aan de examenafname wordt een aantal examenopgaven samen met ankervragen voorgelegd aan een groep personen.
proxy-meting	Indirecte meting van een variabele onder studie, met de verwachting dat deze (sterk) samenhangt met deze variabele.
quick scan	Vier vragen over de kwaliteit van het examen die docenten invullen nadat zij hun scores in Wolf hebben ingevoerd.
referentie-examen	Goed examen uit het recente verleden. De moeilijkheid wordt gebruikt om de norm over te brengen.
schaal	Een reeks getallen die volgens een bepaald voorschrift gekoppeld worden aan waarnemingen.
score	Meestal : het totaal aantal (score)punten dat iemand heeft behaald op een toets. Soms ook : het aantal punten dat op een vraag is behaald.
SE	Schoolexamen.
significant	Aanduiding in de statistiek dat de steekproefuitkomst (met een bepaalde kans) niet aan het toeval kan worden toegeschreven.
standaardbepaling	Activiteit bij centrale examens waarbij meerdere experts (docenten) zich een oordeel vormen over de moeilijkheid van het examen.
standaardfout	Aanduiding in de statistiek voor de precisie van de schatting.
stapelen	Na het behalen van een vo-diploma verder studeren op het naastliggende hogere niveau.
syllabus	Document dat de stof die op het centrale examen bevraagd mag worden, specificeert.
technische N-term	N-term zoals deze uit normhandhavingberekeningen naar voren komt, dus zonder rekening te houden met vakinhoudelijke aspecten.

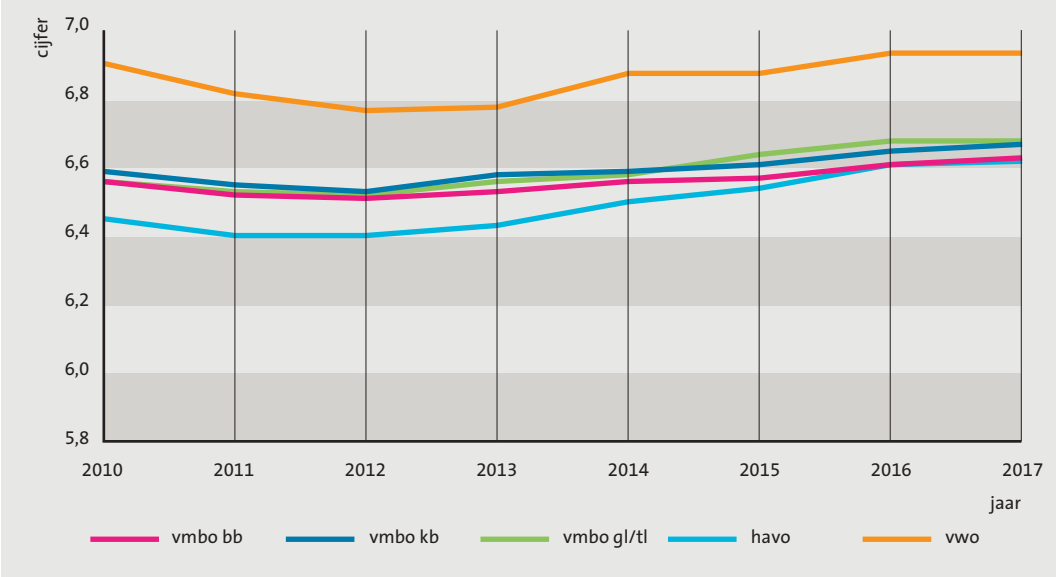
vaardigheid	Bekwaamheid in het uitvoeren van bepaalde handelingen of mentale operaties. Kan ook de beheersing van kennis betreffen.
vmbo bb	Zie leerweg.
vmbo gt	Zie leerweg.
vmbo kb	Zie leerweg.
voorgestructeerde vragen	Vraagtype met structuur waarbinnen het antwoord gegeven moet worden. Zit tussen open vragen en gesloten vragen in.
Wolf	Online applicatie waarmee docenten scores van hun leerlingen (per vraag) aan Cito doorgeven voor de normering.

Bijlagen

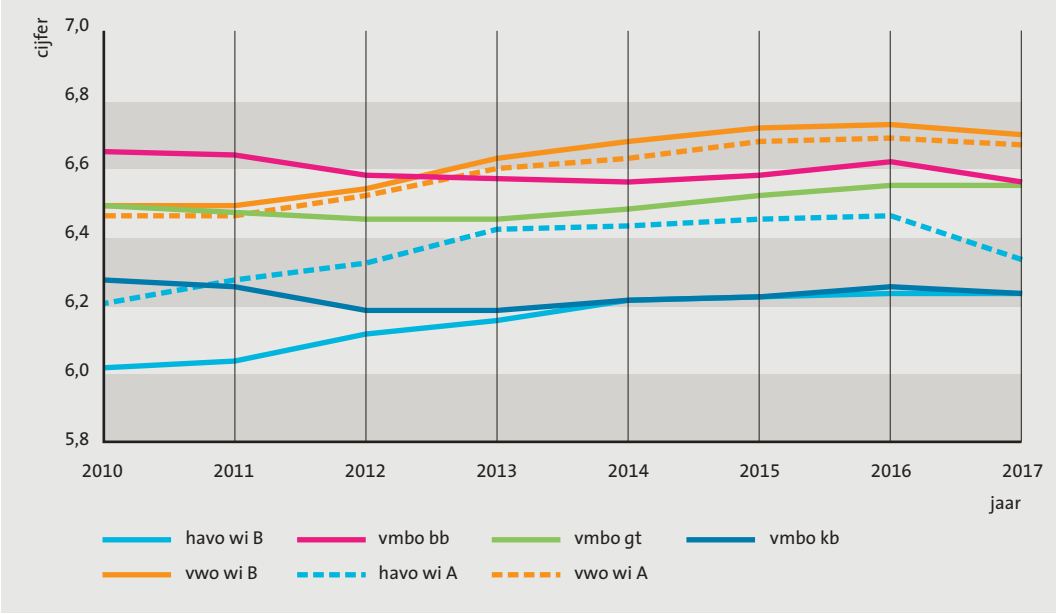
Bijlage 1 | Cijfers schoolexamens in de kernvakken



Figuur 1 Gemiddeld cijfer schoolexamen 2010-2017 Nederlands



Figuur 2 Gemiddeld cijfer schoolexamen 2010-2017 Engels



Figuur 3 Gemiddeld cijfer schoolexamen 2010-2017 wiskunde

Bijlage 2 | De historische ontwikkeling van de normering

De manier waarop we tegenwoordig normeren, is niet altijd hetzelfde geweest. Begrippen als N-term en normhandhaving kennen we pas sinds de eeuwwisseling. Ook heeft de aanscherping van de exameneisen in 2012 en 2013 gezorgd voor wijzigingen in de normeringssystematiek.

1 Vóór de eeuwwisseling: aanname van gelijke moeilijkheid

In de jaren 70 , 80 en 90 werd (behalve bij de moderne vreemde talen) genormeerd volgens het uitgangspunt dat de helft van de score op een examen genoeg was om een voldoende te halen. De examens hadden een lengte van 90 scorepunten. Het cijfer werd als volgt berekend:

Cijfer = score/10 + 1

Een leerling die 45 van de 90 scorepunten behaalde, kreeg dus een 5,5.

Het kwam in die jaren regelmatig voor dat het landelijk gemiddelde voor een vak in drie opeenvolgende jaren fluctueerde. Tabel 1 geeft een voorbeeld voor de vakken aardrijkskunde en wiskunde (mavo-D).

Tabel 1 Landelijk gemiddeld cijfer voor aardrijkskunde en wiskunde mavo-D, in de periode 1998 – 2000.

	1998	1999	2000
Aardrijkskunde	6,3	6,6	5,8
Wiskunde	6,2	6,1	6,8

Lange tijd was dit een geaccepteerde manier van werken. Het uitgangspunt was dat leerlingen een voldoende kregen, wanneer ze de helft of meer van de punten gescoord hadden op een set opgaven, opgesteld door een groep deskundigen en docenten. Impliciet werd er daarbij vanuit gegaan, dat de moeilijkheid van die opgaven in de verschillende examens van jaar tot jaar gelijk was. Werd hier niet aan voldaan en was de fluctuatie groot, dan kon dit leiden tot hogere of lagere exameneisen.

Bekijken we de gemiddelde cijfers in tabel 1, dan zien we duidelijke fluctuaties tussen de verschillende cohorten. Aan de vaardigheid van de leerlingen kan dat niet liggen. Op nationaal niveau is het namelijk niet aannemelijk dat leerlingen het ene jaar aanzienlijk vaardiger zijn dan het andere. Fluctuaties tussen scholen worden uitgemiddeld. Logischer is dat het examen aardrijkskunde in 2000 moeilijker was dan in 1998 en 1999, en dat het examen wiskunde in 2000 juist gemakkelijker was.

Lange tijd kon dit normeringssysteem functioneren. Fluctuaties werden geaccepteerd. Bij de diplomabeslissing werden fluctuaties opgevangen door alle vakken mee te nemen, waardoor moeilijkere en makkelijkere examens elkaar ongeveer ophieven. In de jaren 90 kwam er kritiek op de verschillen in moeilijkheid van jaar tot jaar. Zelfs met gebruik van expertoordelen bleek

het niet mogelijk om de moeilijkheidsgraad van examens voldoende constant te houden. Er moesten andere manieren gevonden worden om meer grip te krijgen op de moeilijkheid van de vragen. Bovendien, de kennis van de moeilijkheid van de opgaven opende de mogelijkheid dat bij de normering rekening kon worden gehouden met de moeilijkheidsgraad van het examen. Halverwege de jaren '90 werden de eerste stappen gezet in 'het project normhandhaving'.

2 De eeuwwisseling: de introductie van normhandhaving

In de jaren '90 startte Cito met het project Normhandhaving. We deden aanvullend onderzoek om vast te stellen of examens een vergelijkbare moeilijkheid hadden als in voorgaande jaren. Deel van het project was dat we de examens van de moderne vreemde talen (die veel gesloten vragen bevatten) ná de examenafname voorlegden aan een vergelijkbare populatie. Het havo-examen werd bijvoorbeeld voorgelegd aan 5-VWO leerlingen. Deze leerlingen maakten stukjes van het actuele examen en stukjes uit een ankerset, en door alles aan elkaar te koppelen kregen wij een beeld hoeveel moeilijker het examen was geweest dan de (geheim gehouden) ankerset. Het jaar daarop deden we dit weer. Als het examen in jaar 1 ongeveer 0,3 cijferpunt moeilijker was dan de ankerset, en het examen van jaar 2 0,7 cijferpunt , dan trokken we de conclusie dat het examen in jaar 2 ongeveer 0,4 cijferpunt moeilijker was dan het examen in jaar 1. Deze werkwijze wordt de posttest genoemd.

We introduceerden ook een andere vorm van onderzoek naar de moeilijkheid van examens: de pretest. Hierbij legden we ruim van tevoren (minstens 2 jaar) examen- en ankeropgaven voor aan leerlingen. Op soortgelijke wijze als bij de posttest kregen we daarmee voorafgaand aan de examenafname een indruk van de moeilijkheid van het examen. (Zie bijlage 2 voor een gedetailleerde beschrijving van het design van pre- en posttesten. Die beschrijft de manier waarop de examens aan elkaar gekoppeld worden via het anker).

Het werken met een pretest levert relatief betrouwbare informatie op. In de praktische uitvoering liggen echter nog wel wat uitdagingen. Zo blijkt soms bij een pretestafname dat een vraag beter anders gesteld kan worden (de vraag was toch niet duidelijk genoeg of de vraag bleek te moeilijk). Gegevens van vragen die nog na de pretest zijn gewijzigd, kunnen we niet meenemen in de pretestberekeningen. Zeer waarschijnlijk is namelijk ook de moeilijkheid van de vraag veranderd.

3 De eeuwwisseling: de introductie van de N-term.

Er zat nog een ander knellend punt in het normeringssysteem van voor de eeuwwisseling. Elk examen (met uitzondering van de examens van de moderne vreemde talen) moest 90 punten bevatten. Vooral bij vakken met veel meerpuntsvragen leidde dit ertoe dat een bepaald type vraag het ene jaar soms 3 en het andere jaar 4 punten waard was. Ook zocht men naar een mogelijkheid om de uitkomsten van de normhandhaving te gebruiken voor een correcte compensatie van de moeilijkheid van een examen .

Om aan deze twee wensen tegemoet te komen, werd de N-term bedacht. Vanaf toen werd het cijfer als volgt bepaald:

Cijfer = score/maximumscore * 9 + N

Er was daarbij nog een klein afhechtingsprobleem. Om ervoor te zorgen dat 0 scorepunten ook altijd het cijfer 1,0 oplevert, en het behalen van alle scorepunten ook altijd resulteert in een 10,0, werden scorepunten aan de bovenkant iets meer of minder waard in termen van cijferpunten.

In de loop der jaren werd de bandbreedte waarbinnen de N-term moest worden vastgesteld, langzaam opgerekt. Inmiddels is deze vastgezet op het interval tussen 0,0 en 2,0.
Als examenmakers streven we naar het maken van een examen met N = 1,0. In dat geval is namelijk de helft van de punten net genoeg voor een voldoende. De gemiddelde N-term over alle vakken heen ligt elk jaar gemiddeld rond de 0,9.

4 De normering in de periode 2000 – 2011

In de jaren 90 werd de normhandhavingsprocedure verder ontwikkeld. Vanaf 2000 spelen de uitkomsten een rol bij de normering. In deze periode werd tot 2011 een budget ter beschikking gesteld voor normhandhavingsonderzoeken. Omdat dit niet voldoende was om voor alle vakken een pretest of posttest te organiseren, moesten we keuzes maken. Die keuzes werden ingegeven door de uitvoerbaarheid, het belang van een vak en hoeveel leerlingen het vak gewoonlijk kiezen.

De keuze voor een post- of pretest bepaalden we als volgt. Een posttest is alleen goed uitvoerbaar bij vakken met veel gesloten vragen. Want omdat een posttest pas na de examens op scholen wordt afgenomen (in verband met de geheimhouding kan het niet eerder) en de resultaten bekend moeten zijn ten tijde van de normering (half juni), is de verwerkingstijd kort. Te kort om een handmatige correctie van open vragen te doen. Om die reden wordt een posttest vooral ingezet bij de moderne vreemde talen. De enige uitzondering is Frans vwo. Leerlingen in 5 vwo zijn namelijk niet vaardig genoeg om het vwo-examen fatsoenlijk te kunnen maken.

Een vak is geschikt voor een pretest als de inhoud stabiel is. De afname van examenvragen twee jaar voor dato levert immers alleen zinvolle informatie op, wanneer leerlingen op dezelfde wijze over dezelfde stof les hebben gehad, als leerlingen die twee jaar later examen doen. Voor bijvoorbeeld geschiedenis bleek deze randvoorwaarde knellend, maar bij de exacte vakken vormde het geen probleem. Hoewel ook bij de mens en maatschappijvakken en de kunstvakken sindsdien geprobeerd is om normhandhaving van de grond te krijgen, had dat weinig succes. Eigenlijk is alleen economie vmbo gl/tl een succesvolle pretest geworden.

Voor vakken met een posttest of pretest bepaalden we de moeilijkheidsgraad van nieuwe examens via een vergelijking met het referentie-examen. Dit is een eerder afgenomen examen dat goed functioneerde, de juiste moeilijkheidsgraad had en waarin geen calamiteiten of onduidelijkheden zaten. Het referentie-examen dient als voorbeeldexamen en ijkpunt.
Zodra bekend is hoe de moeilijkheid van een examen zich verhoudt tot het referentie-examen, kunnen de waargenomen scores worden geduid. Hoge scores kunnen bijvoorbeeld veroorzaakt worden door een makkelijk examen óf door goede leerlingen. Als we weten dat het huidige examen net zo moeilijk is als het referentie-examen en de huidige scores liggen hoger, ligt het voor de hand te concluderen dat de huidige leerlingen beter zijn.

Voor vakken zonder posttest of pretest konden we de moeilijkheidsgraad van nieuwe examens bepalen via een aanname. Die aanname was dat de vaardigheid van de totale examenpopulatie van jaar tot jaar weinig verandert. Door een referentie-examen aan te wijzen, kon met de aanname van gelijke populaties een N-term worden gevonden waarbij het percentage onvoldoendes en het gemiddelde cijfer van het huidige examen gelijk was aan die van het referentie-examen. Met andere woorden: bij deze vakken werd in die tijd relatief genormeerd.

De aanname van gelijke populaties (even vaardige populaties over jaren heen) werd in 2010 in twijfel getrokken. Er was in het vmbo sprake van een opstroom. Steeds minder leerlingen deden vmbo bb-examen en het aantal vmbo gl/tl-kandidaten nam toe. We deden vervolgens in retrospectief onderzoek om na te gaan of leerlingen in de loop van de tijd minder vaardig

geworden waren. Hiertoe vergeleken we de examens 2005, 2007 en 2009 met elkaar. De conclusie, die in de loop van 2010 aan het licht kwam, was een lichte daling van de vaardigheid van vmbo-leerlingen. In 2011 werd de normering van de vmbo-examens met deze informatie in lijn gebracht.

5 De opkomst van de digitale examens

Sinds 2005 worden in het vmbo digitale examens afgenomen. Eerst alleen voor algemene vakken op het niveau vmbo bb, een paar jaar later ook voor het niveau vmbo kb. Bij een digitale examinering zijn van een examen meerdere varianten beschikbaar, die in de periode van april tot en met juni kunnen worden afgenomen. In tegenstelling tot de reguliere examens blijven deze examens geheim na de afname. Dit biedt de mogelijkheid voor een andere normerings-techniek.

Bij digitale examens kunnen dezelfde opgaven worden voorgelegd aan verschillende cohorten leerlingen. Hiermee kan worden gemeten of de huidige examenpopulatie beter of slechter presteert dan eerdere populaties. Deze vorm van normeren wordt 'Anchor in Package' (AIP) genoemd. Het voordeel van AIP is dat alle examenkandidaten ankeropgaven maken; deze ankervragen maken integraal onderdeel uit van het examen. Voordelen zijn dat de ankervragen door meer leerlingen worden gemaakt dan bij een pre- of posttest. Doordat de ankeropgaven bij AIP onderdeel zijn van het eigenlijke examen, zijn leerlingen bovendien gemotiveerder om goed te presteren. En doordat ankeropgaven enkele jaren later opnieuw ingezet kunnen worden, maken verschillende cohorten exact dezelfde opgaven onder dezelfde condities. Het gevolg is dat AIP bijzonder nauwkeurige schattingen geeft van de moeilijkheid van een examen en de vaardigheid van een populatie.

De digitale examens voorzagen in een behoefte, getuige het feit dat binnen een aantal jaar vrijwel alle scholen in Nederland overstapten . Tabel 2 laat zien hoeveel leerlingen digitaal examen deden in de algemene vakken (vmbo bb en kb, 2006 – 2013). Nog steeds is een digitale afname niet verplicht. Dat betekent dat scholen nog steeds kunnen kiezen voor papieren examens. Dit zijn examens met deels een andere inhoud en – omdat ze de laatste jaren nog door zeer weinig leerlingen worden gemaakt – een andere normering. Daarbij wordt verondersteld dat de leerlingen die de papieren examens maken even vaardig zijn als de leerlingen die de digitale examens maken.

Tabel 2 Percentage kandidaten dat digitaal examen doet in de algemene vakken in vmbo bb en kb

	2006	2007	2008	2009	2010	2011	2012	2013
bb	18	42	86	94	>99	>99	>99	>99
kb	-	-	-	-	11	40	72	82

6 Wijziging zak/slaagregeling: De normering met Fisher 2012 – nu

In 2012 en 2013 werd de zak/slaagregeling aangepast. In 2012 werd de CE-eis ingesteld: leerlingen moesten minimaal een 5,50 halen voor het gemiddelde van hun centrale examens en in 2013 volgde de kernvakkenregel. Havo/vwo-leerlingen mochten voortaan maximaal één vijf als eindcijfer scoren voor de kernvakken. Op het vmbo betekende de kernvakkenregel dat met ingang van 2014 voor Nederlands minimaal een vijf gehaald moest worden. Bij gelijkblijvende

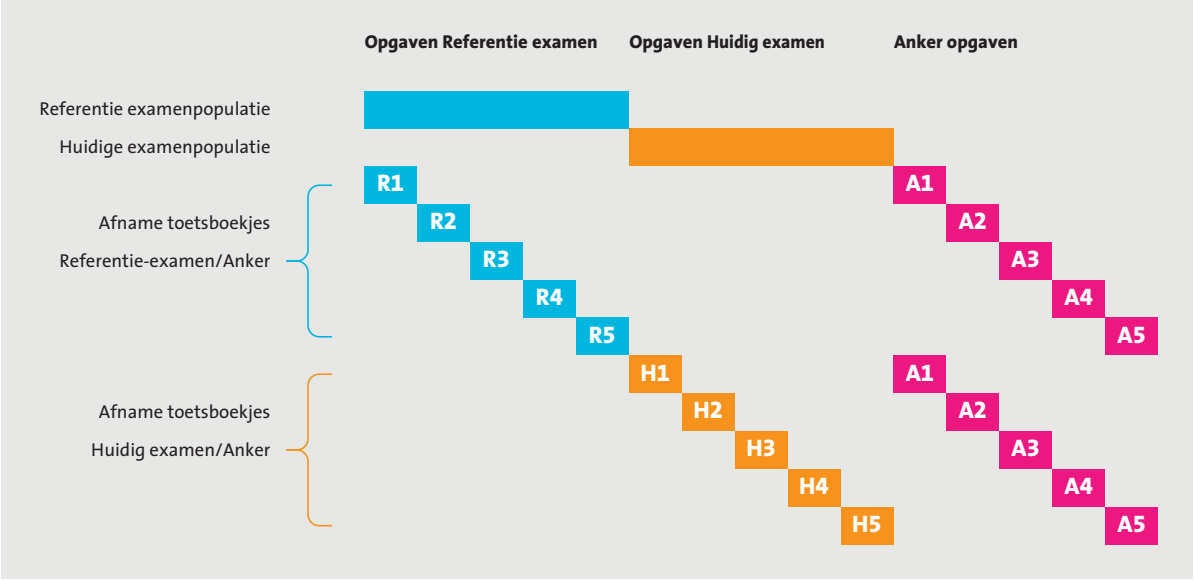
prestaties (als in 2010 en 2011) zouden al deze aanpassingen leiden tot een groter aantal gezakte leerlingen. Verwacht werd dat scholen en leerlingen zouden inspelen op de nieuwe eisen met betere prestaties op de centrale examens als gevolg. De aanname van gelijke populaties kon daardoor niet langer worden gebruikt bij de normering.

In 2012 werd de normeringsprocedure voor vakken zonder aanvullende gegevens (zoals een pretest en posttest) aangepast. Voor 2012 werden deze vakken genormeerd onder aanname van gelijkblijvende populaties. Sinds 2012

1) wordt aangenomen dat een vaardigheidsverandering bij vakken zonder aanvullende normeringsgegevens vergelijkbaar is met een vaardigheidsverandering bij vakken waarvoor wel extra normhandhavingsgegevens zijn verzameld. Daarbij worden vakken die inhoudelijk en procedureel op elkaar lijken (bijvoorbeeld kernvak en niet-kernvak), geclusterd;
2) worden bij een aantal vakken standaardbepalingen ingezet. Hierbij bepaalt een groep experts de moeilijkheid van de examens. Dergelijke standaardbepalingen zorgen voor aanvullende gegevens. De standaardbepaling is veel minder nauwkeurig dan norm-handhavingsonderzoeken via pretest, posttest of AIP. De inschatting van de moeilijkheid van opgaven fluctueert namelijk tussen beoordelaars, waardoor de moeilijkheid van examens niet nauwkeurig te schatten is en daarmee ook de schatting van de vaardigheid van de populatie onnauwkeurig is. Dat neemt niet weg dat via de extra standaardbepaling waardevolle gegevens over vaardigheidsontwikkelingen naar boven worden gehaald. Gegevens die anders voor bijvoorbeeld mens en maatschappijvakken of kunstvakken niet beschikbaar zouden zijn.

Bijlage 3 | Designs bij pre- en posttest

Bij zowel pre- als posttest worden opgaven voorgelegd aan leerlingen in een boekje. Een boekje is hierbij een 'toets' waarin examenopgaven zijn gemengd met ankeropgaven. Dat betekent dat de leerlingen in één zitting zowel examenopgaven maken als ankeropgaven. De variabiliteit bij post- en pretesten worden in sterke mate bepaald door de kracht van het design. Met het design wordt bedoeld de overlap tussen de boekjes en het aantal leerlingen dat die boekjes heeft gemaakt.
Wanneer een recent examen gelinkt moet worden aan een referentie-examen gaat dit via het anker. De kracht van de link tussen huidig en referentie-examen is dus afhankelijk van de kracht van het design tijdens de pretest van het huidige examen en die van de pretest ten tijde van het referentie-examen.
Bij posttests worden alleen de gesloten vragen meegenomen. Hoe groter het aandeel gesloten vragen hoe beter de link, hoe kleiner de standard error.
De opzet van een pre- en posttestdesign is weergegeven in de figuur hieronder.
Hierin is te zien dat de opgaven uit het referentie-examen (blauw) verdeeld worden over meerdere boekjes en vervolgens aan leerlingen worden voorgelegd samen met ankeropgaven. Hiermee wordt informatie verkregen over de moeilijkheid van de opgaven van het referentie-examen ten opzichte van de ankervragen. Het huidige examen (oranje) wordt ook samen met dezelfde ankervragen voorgelegd aan leerlingen. Hiermee wordt informatie verkregen over de moeilijkheid van het huidige examen ten opzichte van het anker. Met deze werkwijze wordt het mogelijk om uitspraken te doen over de moeilijkheid van het huidige examen ten opzichte van het referentie-examen.



Bijlage 4 | Onvolkomen vragen en procedures daaromheen

Elk jaar bevatten de centrale examens duizenden vragen. En elk jaar zitten er een paar fouten in waarvan je kan zeggen: 'Hoe hebben ze dat nou over het hoofd kunnen zien?' Het examen wordt door een handjevol mensen gescreend. Niet te veel want dat brengt de geheimhouding in gevaar en niet te weinig in verband met de kwaliteit. De screeners komen uit het vervolgonderwijs of hebben anderszins een reputatie opgebouwd als vakexpert. Zij beoordelen een tussenproduct op vakinhoudelijke juistheid en kwaliteit van vraagstelling. Het komt voor dat alle screeners een fout over het hoofd zien, maar het komt ook voor dat er last minute (kort voor aanlevering aan de drukker) nog iets aan het examen veranderd wordt, waarbij er onbedoeld een onvolkomenheid in sluipt.

In de examencampagne liggen de examens onder een vergrootglas. Dat is goed want voor de leerlingen hangt er veel van af. Zo ontstaat elk jaar discussie over de examenvragen zelf, over het antwoordmodel of over de combinatie van die twee. In een aantal gevallen leidt de discussie tot de vraag of het niet beter is om de vraag te schrappen. Zo'n discussie kan veroorzaakt worden doordat de vraag op meerdere manieren geïnterpreteerd kon worden. Het kan ook komen doordat men van mening is dat een ander antwoord ook juist is of omdat men de inhoudelijke juistheid van het gegeven antwoord in twijfel trekt. Het is daarbij vaak geen zwart-wit-situatie. Is aujourd'hui één of twee woorden? Is √8 ook goed als 2√2 het antwoord moet zijn? Rekkelijken en preciezen staan lijnrecht tegenover elkaar en kunnen het soms maar moeilijk eens worden met elkaar.
De examenlijn is de plek waar docenten kunnen melden dat ze vinden dat een vraag niet deugt. Voordat een beslissing wordt genomen of de vraag gehandhaafd kan blijven, zijn er veel mensen die hierover meedenken. Dit zijn natuurlijk de leden van de vaststellingscommissie (docenten) en de toetsdeskundige van Cito, maar ook andere mensen (native speakers, vakexperts etc).
Dit artikel wil laten zien dat de discussies over het al dan niet handhaven van een vraag soms heel subtiel kunnen zijn.

Moment van besluit of een vraag onvolkomen is

Er zijn twee mogelijkheden. Het besluit dat een vraag onvolkomen is, wordt genomen binnen vier werkdagen na afloop van het examen, of het besluit valt daarna[21]. In het eerste geval kan een inhoudelijke aanvulling op het correctievoorschrift worden uitgedaan. Met een aanpassing op het correctievoorschrift kan het probleem vaak worden verholpen.
Als het besluit ná deze vier werkdagen valt, wordt er normaal gesproken geen aanvulling op het correctievoorschrift meer uitgedaan. Mededelingen over het neutraliseren van een vraag kunnen nog wel tot iets meer dan vier werkdagen gepubliceerd worden. Na afloop van de vier werkdagen worden alle klachten over ondeugdelijkheden bewaard tot de normeringsvergadering. Op de normeringsvergadering worden deze klachten uitgebreid besproken en kan alsnog worden besloten dat een vraag onvolkomen is. Als dat het geval is, wordt berekend wat de hoogte van de ophoging van de N-term moet zijn. Deze ophoging is bedoeld om ervoor te zorgen dat leerlingen die last hadden van de onvolkomenheid, toch het cijfer krijgen dat ze

21 In bijzondere gevallen kan van de deadline van vier werkdagen worden afgeweken.

verdienen. Een nadeel van deze werkwijze is dat er ook leerlingen zijn die door de compensatie een hoger cijfer krijgen dan ze verdienen. Dit is onwenselijk. Naast het feit dat dit oneerlijk voelt tegenover de leerlingen van andere jaren, voelt het ook oneerlijk voor de leerlingen die wél last hadden van de onvolkomenheid. Bovendien mag je stellen dat de waarde van het diploma daalt wanneer veel leerlingen een hoger cijfer krijgen dan ze verdienen. Het uitgangspunt bij een ophoging van de n-term voor een onvolkomen vraag is dat de kandidaat die geen punten heeft kunnen scoren op de onvolkomen vraag, precies voldoende wordt gecompenseerd.

De periode van vier werkdagen

De periode van vier werkdagen is historisch zo gegroeid. Hoe korter de periode, hoe minder aanvullingen er uitgedaan worden. Minder aanvullingen zorgt dan misschien voor minder duidelijkheid en veelal tot meer discussie tussen eerste en tweede corrector. Vaak is het zo dat de aanvulling duidelijkheid geeft waardoor de eerste en tweede corrector geen discussie meer hoeven te voeren. Hoe langer de periode, hoe meer aanvullingen, hoe beter, zou je dus zeggen. Maar heel erg late aanvullingen leiden ook tot frustraties. Eerste correctoren die al klaar waren met corrigeren moeten bij een late publicatie van een aanvulling alles weer opnieuw bekijken. Daarbij komt dat het werk soms al naar de tweede corrector is gestuurd en de eerste corrector de aanpassing niet meer op het leerlingwerk kan verwerken. Dit alles leidt tot veel administratief gedoe. Eindconclusie: op dit moment wordt de termijn van vier werkdagen als een redelijk compromis ervaren.

Het neutraliseren van een vraag

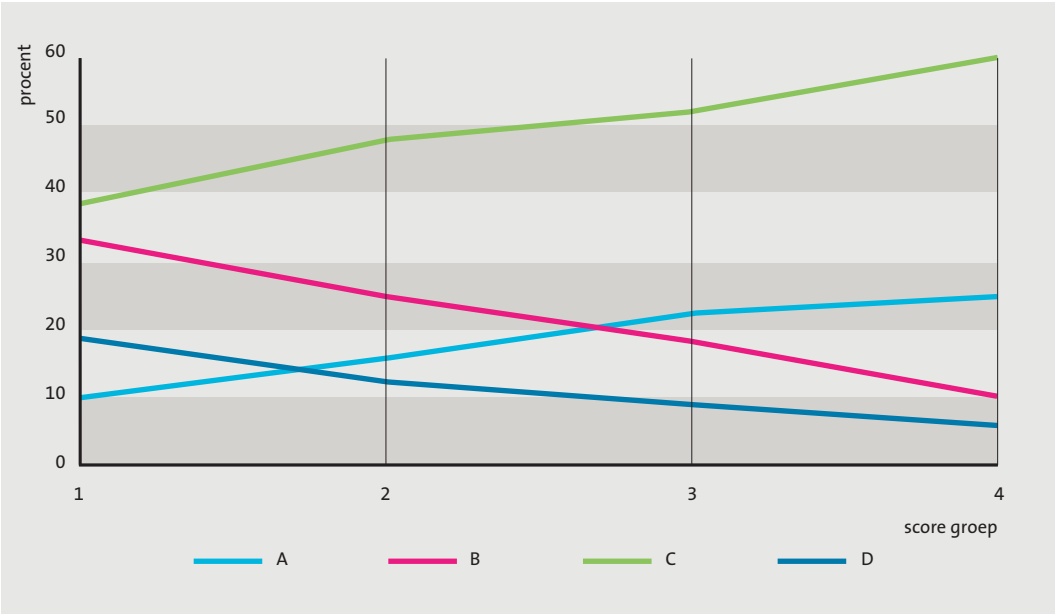
Een bijzondere situatie vormen de vragen waarbij het niet mogelijk is om met een ‘gewone’ inhoudelijke aanvulling op het correctievoorschrift het probleem te verhelpen. Er rest dan eigenlijk nog maar één mogelijkheid: de vraag weg te geven. In jargon: te neutraliseren.
Dit betekent dat alle leerlingen de maximale score krijgen voor die vraag, ongeacht het gegeven antwoord. Er zitten echter ook nadelen aan het neutraliseren van een vraag: Ten eerste zullen leerlingen die de vraag goed hadden, het gevoel krijgen dat het niet eerlijk is: ‘nu krijgt iedereen de volle score voor die vraag!’ Ten tweede hebben alle leerlingen aandacht en tijd aan de opgave besteed. Ten derde blijkt vaak (niet altijd) uit de toetsanalyse dat de vraag het prima doet. Daarmee bedoelen we dat de vraag een goed onderscheid maakt tussen goede en minder goede leerlingen. Het neutraliseren van de vraag zorgt dus voor een minder onderscheidend examen. Een korte toelichting: Na het neutraliseren van een vraag krijgt de groep minder goede leerlingen er meer punten bij dan de groep met betere leerlingen. De spreiding van de totaalscores wordt minder groot en het examen maakt dan minder goed onderscheid tussen goede en minder goede leerlingen.
Het is dus niet verstandig om al te lichtvoetig een vraag te neutraliseren. Daarmee gooi je nuttige en bruikbare informatie overboord. De redenen om een vraag te neutraliseren moeten dus onomstotelijk en overtuigend zijn.

Twee voorbeelden

Veel mensen zijn van mening dat een examenvraag deugt of niet deugt. Toch komt het voor dat een vraag soms half deugt. Een vraag roept soms discussies op waarbij men niet gemakkelijk tot overeenstemming komt over wat er gedaan moet worden. Onderstaande voorbeelden zijn bedoeld om te laten zien dat een onvolkomen vraag niet per se een fout hoeft te bevatten. We spreken daarom ook liever van een onvolkomen vraag dan van een foute vraag. Het gaat namelijk vaak om een verschil van inzicht tussen docenten over hoe om te gaan met (vaak onverwacht) antwoordgedrag van leerlingen.

Engels vwo

In het examen Engels vwo van 2017 ontstond discussie over vraag 18. Deze vraag bij de tekst 'Britain's Brussels syndrome' vroeg welke uitspraak geldig was, volgens paragraaf 4 en 5. De kritiek was dat niet iedereen vond dat het goede antwoord overduidelijk in de tekst stond vermeld. Tijdens de analyse van de afnamegegevens bleek dat er iets vreemds met deze vraag aan de hand was. Om dit te illustreren is onderstaande grafiek weergegeven. Een korte uitleg: in de grafiek staan op de horizontale as de vier kwartielen van leerlingen. Het meest linker punt correspondeert met de 25% zwakste kandidaten en het meest rechtse punt met de 25% beste, meest vaardige kandidaten. De vier lijnen corresponderen met de vier antwoorden. De groene lijn (C) correspondeert met het goede antwoord. De lijn laat zien hoeveel procent van de vier kwartielen leerlingen voor dit antwoord gekozen hebben. Bij een goed onderscheidende vraag kiest de betere leerling vaker voor het goede antwoord. Bij het goede antwoord hoort dus een stijgende lijn. Bovendien zou het zo moeten zijn dat de groep betere leerlingen minder vaak kiest voor een verkeerd antwoord. Bij de verkeerde antwoorden horen dus dalende lijnen. Wat opvalt in onderstaande grafiek is dat een van de verkeerde antwoorden (de lichtblauwe lijn) een stijgende lijn laat zien. De betere leerlingen kiezen dus vaker voor dit verkeerde antwoord dan de minder goede leerlingen. Deze informatie, samen met de inhoudelijke discussie over de vraag, heeft ertoe geleid dat op de normeringsvergadering alsnog is besloten deze vraag onvolkomen te verklaren.



M&O havo

In het M&O examen 2017 werd gesteld dat een bedrijf in 2018 voor 1500 euro haar inventaris wilde vervangen. De vraag was of er volgens het gegeven overzicht over 2017 hiervoor voldoende geld (liquide middelen) aanwezig zou zijn. Bij het antwoord volgens het correctie-voorschrift konden antwoorden van vorige vragen worden gebruikt om te concluderen dat er voldoende geld voorhanden zou zijn. Tot zover geen probleem. Nu waren er leerlingen die op een andere manier aan de slag gingen. Zij gingen een balans opstellen met de liquide middelen als sluitpost. Daarbij veronderstelden deze leerlingen dat alle relevante gegevens aanwezig waren. Dat was niet zo en daarmee gingen ze de mist in. Toch was hun gedachte niet verkeerd en de aanname dat alle gegevens aanwezig waren, was ook best verdedigbaar. De vak-vereniging besloot na de examenbespreking tot het goedkeuren van deze werkwijze. Toch kwamen er geluiden binnen dat er nog steeds docenten waren die het antwoord volgens deze alternatieve methode fout rekenden. Om die reden werd de vraag alsnog onvolkomen verklaard.

Bijlage 5 | Populatieschattingen met plausible values

Introductie

Waar examens primair de vaardigheid van een individuele leerling in kaart brengen, schatten onderwijskundige surveys als PISA de vaardigheid van een populatie. Traditionele methoden zoals *marginal maximum likelihood* (MML) en *expected-a-posteriori* (EAP) zijn puntschattingen optimaal voor het schatten van vaardigheden van individuele leerlingen, maar minder geschikt voor populatieschattingen. Worden ze in zo'n geval toch gebruikt, dan kunnen deze de resultaten vertekenen.

Een manier om toch tot zuivere populatieschattingen te komen, kan door voor elke leerling meerdere waarden die de verdeling van de vaardigheid van een leerling weergeven te gebruiken. Deze zogenaamde *plausible values* geven per leerling een set van vaardigheidsscores. Die kunnen vervolgens ingezet worden voor het zuiver schatten van de populatievaardigheid (Von Davier, Gonzalez & Mislevy, 2009).

Plausible values werden voor het eerst gebruikt in de analyse van de *National Assessment of Educational Progress* (NAEP; Mislevy et al, 1992). Ze zijn gebaseerd op de theorie van *multiple imputation*, ontwikkeld door Rubin (Rubin, 1987) en worden sindsdien in steeds meer (internationale) peilingsonderzoeken gebruikt.

Het werken met plausible values is voor onderwijskundige surveys ook om andere redenen interessant. Door een individuele vaardigheid in kaart te brengen met meerdere waarden, wordt de onzekerheid van de meting verdisconteerd. Voor onderwijskundige surveys is dit relevant, omdat vaak gebruik wordt gemaakt van relatief korte toetsen, waarbij de vaardigheid van leerlingen niet nauwkeurig geschat kan worden.

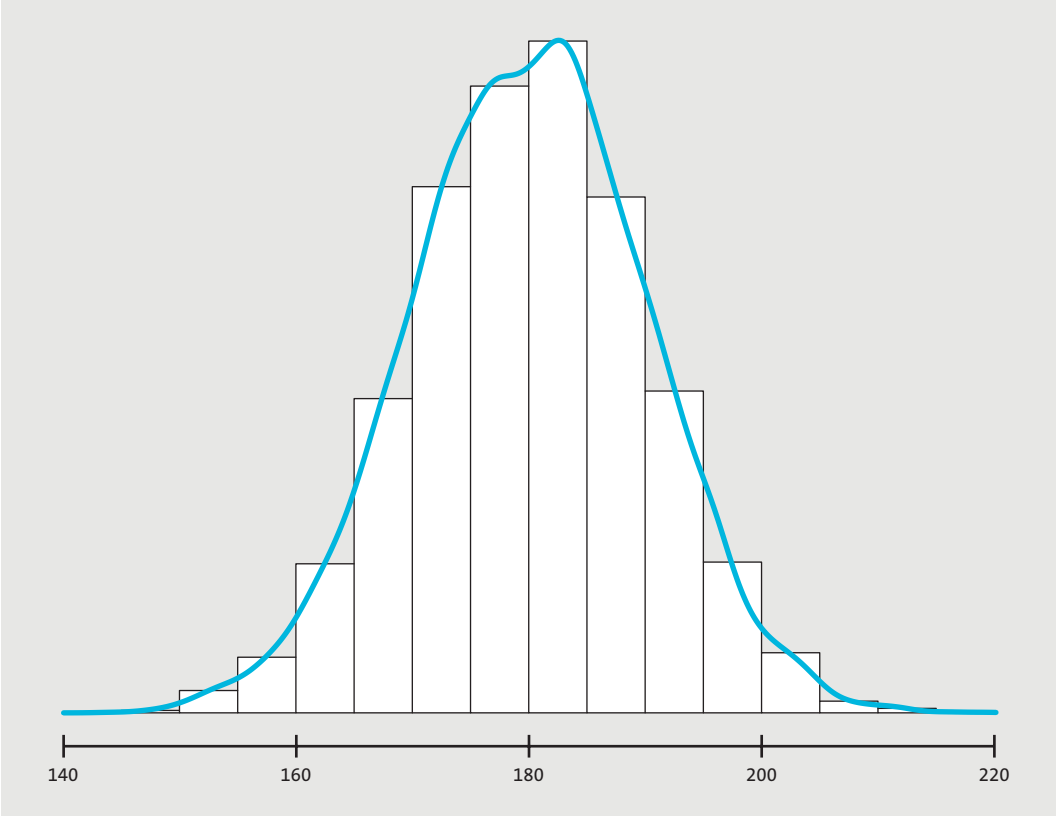
Naast dat leerlingen in onderwijskundige surveys vaak relatief korte toetsen maken, maken ze ook vaak verschillende toetsen. Via deze verschillende toetsversies kan bij korte toetsen dan toch het hele domein van opgaven dat een construct meet, worden waargenomen. De scores op de verschillende toetsversies moeten dan wel eerst geëquivaleerd worden. Dit gebeurt met een IRT-model dat scores corrigeert voor de moeilijkheid van de toets. Plausible values kunnen daarmee beschouwd worden als de geëquivaleerde scores op het hele domein, op basis van item parameters uit een IRT kalibratie. Het zuiver schatten van deze item parameters is een cruciale stap in de berekening van plausible values.

Deze bijlage beschrijft de achterliggende principes van de plausible values-methode. Bij gebruik van verschillende toetsversies is het essentieel dat verschillen in moeilijkheidsgraad tussen toetsen juist verdisconteerd worden. Hierin spelen in de procedure die PISA gebruikt item parameters uit een IRT model een belangrijke rol. De problemen bij het schatten van item parameters - nodig voor het schatten van plausible values - worden in het tweede gedeelte van deze notitie besproken.

Plausible values en populatieschattingen

Om te illustreren hoe de plausible values-methodologie werkt, geven we een voorbeeld. Hierin zijn 10 000 observaties gesimuleerd, met een gemiddelde van 180 en een

standaarddeviatie van 10. Het zou bijvoorbeeld kunnen gaan om de lengte van personen. We beschouwen deze gegevens als de ware, maar onbekende populatiewaarden. De verdeling van de populatieverdeling ziet er als volgt uit:



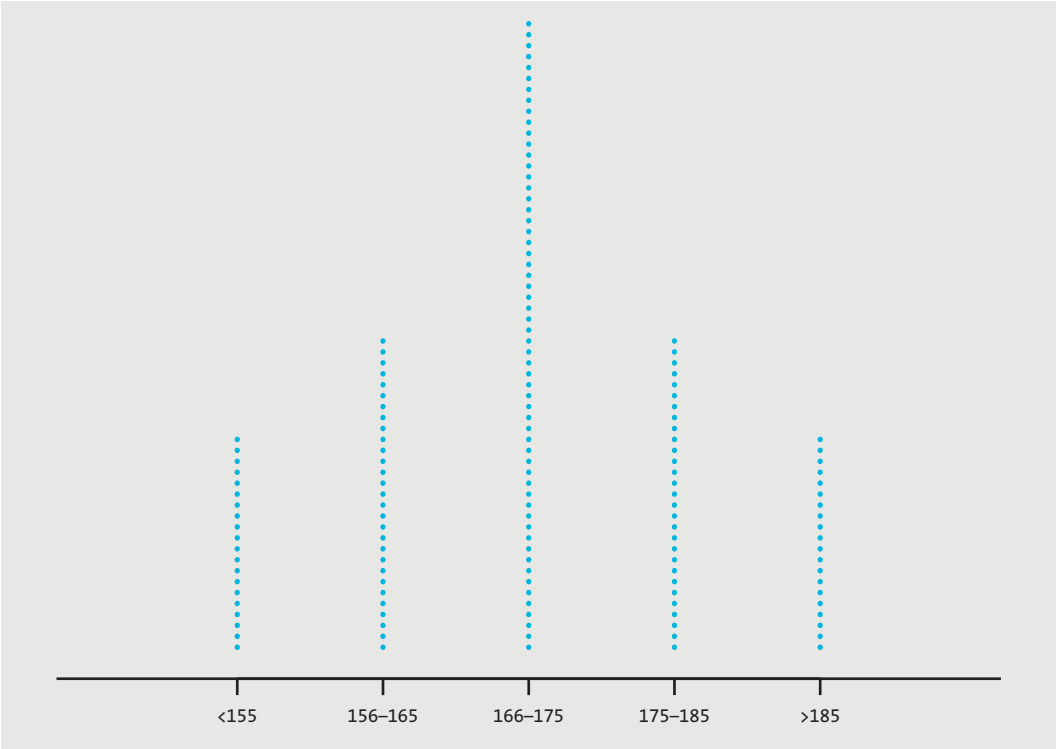
Figuur 1 Histogram van de populatieverdeling

Stel nu dat we de lengte van mensen in kaart willen brengen met behulp van een vragenlijst. Een grove manier om lengte te meten, is met een gesloten vraag voorzien van slechts enkele (hier vijf) antwoordcategorieën. Dit levert de volgende informatie op voor de eerste vijf gesimuleerde personen:

Tabel 1 Ware lengte en gerapporteerde categorie

Persoon	Ware lengte	Gerapporteerde categorie	Label gerapporteerde categorie
1	185,08	5	>185
2	180,66	4	175-185
3	179,61	4	175-185
4	199,85	5	>185
5	174,62	3	166-175

De verdeling met geobserveerde waarnemingen ziet er nu als volgt uit:



Figuur 2 Verdeling van observaties verkregen met behulp van een vragenlijst

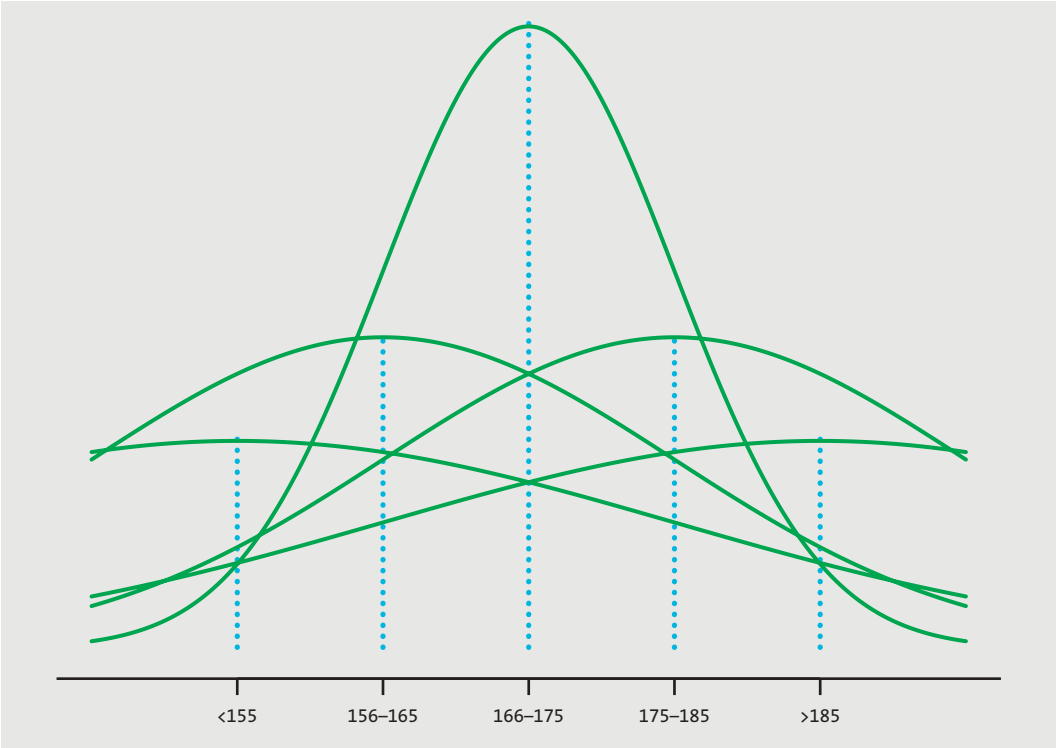
In tegenstelling tot de ware populatieverdeling is de geobserveerde verdeling discreet: mensen konden hun lengte met maar vijf antwoordcategorieën aangeven. Uiteraard is dit een extreem voorbeeld op basis van slechts één vraag. Bij een volledige toets is het principe echter hetzelfde: het aantal mogelijke geobserveerde scores is beperkt.

Het beperkte aantal scoremogelijkheden geeft niet noodzakelijkerwijze de (continue) populatieverdeling accuraat weer. Plausible values bieden dan een oplossing voor zowel het beperkt aantal mogelijke scores, als het verdisconteren van (verschillen in) meetonzekerheid. Bij de berekening van plausible values moeten in principe twee stappen worden uitgevoerd:

1) De berekening van een verdeling rondom de geobserveerde score. Deze verdeling wordt de posterior-verdeling genoemd [22].
2) Het trekken van een set van toevallige waarden uit deze posterior-verdeling voor iedere leerling (Wu, 2005).

Normaal gesproken zijn vaardigheidsschattingen van leerlingen met meer extreme scores minder betrouwbaar dan vaardigheidsschattingen van leerlingen met moderate scores. Dit reflecteert de grotere meetonzekerheid aan de uiteinden van de verdeling. In het voorbeeld zou de posterior-verdelingen rondom de gerapporteerde waarden er dan ook als volgt uit kunnen zien:

22 De posterior-verdeling is een waarschijnlijkheidsverdeling die weergeeft wat je weet nadat je de data hebt gezien.



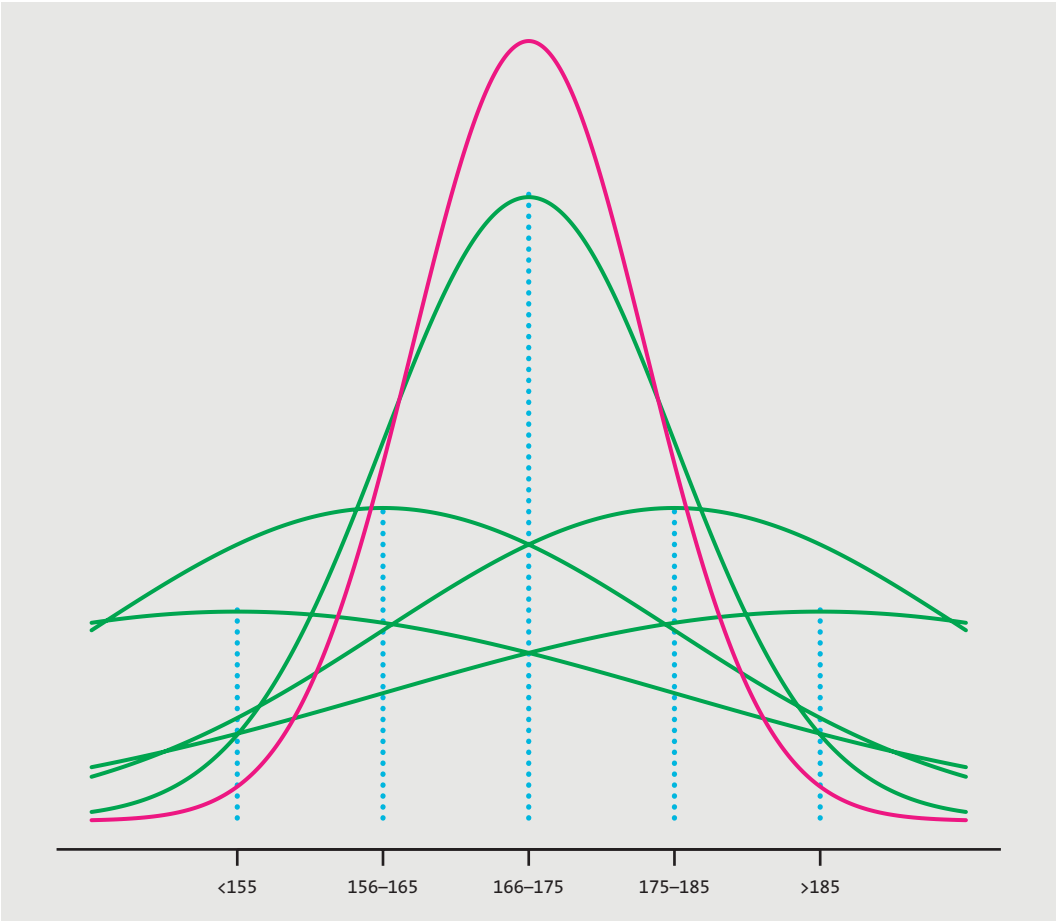
Figuur 3 Posterior-verdelingen rondom de gerapporteerde waarden

Na het maken van de posterior-verdelingen wordt dan een set van random getallen getrokken uit de verdelingen. Om het voorbeeld eenvoudig te houden, trekken we hier één random getal. De data ziet er dan als volgt uit:

Tabel 2 Ware lengte, gerapporteerde lengte en plausible value

Persoon	Ware lengte	Gerapporteerde categorie	Label gerapporteerde categorie	PV1
1	185,08	5	>185	1.956637
2	180,66	4	175-185	5.777844
3	179,61	4	175-185	2.549571
4	199,85	5	>185	4.341423
5	174,62	3	166-175	3.059529

Daarbij bevat kolom 2 de ware en onbekende waarde, en tonen kolom 3 en 5 de gerapporteerde waarde en de eerste plausible value. Dit maakt direct duidelijk waarom plausible values niet geschikt zijn voor individuele scoring: mensen met dezelfde gerapporteerde scores kunnen een totaal andere plausible value toegewezen krijgen. Voor het schatten van populatieverdelingen zijn plausible values echter zeer geschikt. Dat blijkt uit de volgende figuur, waarin de dikkere rode lijn de plausible values-verdeling weergeeft.



Figuur 4 Verdeling van plausible values

Wat blijkt is dat de verdeling van de plausible values, veel meer dan de verdeling van de gerapporteerde waarden, de verdeling van de populatie volgt.

Schatting van item parameters

In PISA wordt gebruik gemaakt van meerdere toetsversies. Daardoor kan de geobserveerde score niet zomaar worden gebruikt als startpunt voor het construeren van een posterior-verdeling. Het aantal goed op de ene toetsversie vereist namelijk niet noodzakelijkerwijs dezelfde vaardigheid als hetzelfde aantal goed op een andere toetsversie. Hiervoor moeten de toetsversies eerst geëquivaleerd worden.

Het correct schatten van item parameters is - zoals eerder aangegeven - een cruciale stap in het vergelijkbaar maken van toetsversies, en daarmee in de berekening van plausible values. PISA heeft - net zoals andere internationale onderwijskundige surveys - de ambitie om zowel vergelijkingen te maken tussen landen als tussen afnames in verschillende jaren. Voor items die in meerdere afnames voorkomen, wordt in eerste instantie geprobeerd de moeilijkheid ervan in kaart te brengen met behulp van dezelfde parameter. Lukt dit niet, dan wordt de moeilijkheid geschat met twee of meer afzonderlijke item parameters. Zo wordt dan net gedaan alsof het twee verschillende items zijn. Om vergelijkingen tussen afnames te kunnen blijven maken, moeten in deze procedure wel voldoende items overblijven die in alle afnames met dezelfde item parameter valide zijn geschat.

Het vergelijkbaar maken van toetsresultaten met een IRT-kalibratie is zeer gangbaar in allerlei vergelijkingsstudies en niet beperkt tot het gebruik van plausible values. De ambities van internationale studies zoals PISA zijn echter veel uitgebreider dan vele andere studies. In de methodologie van PISA vraagt deze ambitie bijvoorbeeld dat de moeilijkheid van items afgenomen in Costa Rica in 2015, met dezelfde parameter geschat kan worden als dezelfde items afgenomen in Nederland in 2012. Vandaar dat juist rondom internationale surveys veel wetenschappelijke discussie bestaat rondom *Differential Item Functioning* en *Measurement invariance* (zie bijvoorbeeld Davidov, Schmidt, Billiet & Meuleman, 2018).

De schatting van de item parameters - en daarmee de trendvergelijkingen - tussen PISA-2012 en 2015 staan echter ter discussie. In 2015 stapte PISA over op een digitale toetsafname. De trendvergelijkingen werden toen gebaseerd op de aanname dat items afgenomen op papier (in PISA-2012) even moeilijk waren als dezelfde items afgenomen op een computer (in PISA-2015). Deze aanname werd voorafgaand aan de afname van PISA-2015 onderzocht in een *field trial*. Wat volgde was een set items die niet onderhevig waren aan verschillen tussen afnamemodes. Na de afname in 2015 werden daarvan echter toch enkele items apart geschaald (OECD, 2017). Inmiddels is echter duidelijk dat deze correctie voor mode-verschillen waarschijnlijk niet voldoende is geweest omdat niet alle landen op gelijke wijze reageren op een andere afnamemode. Ook Andreas Schleicher, hoofd van de PISA-studie, heeft aangeven dat de trendvergelijkingen in 2015 onder druk staan: *"...country-by-mode differences require further investigation – to understand whether they reflect differences in computer familiarity, or different effort put into a paper test compared to a computer test* [23]*"*. Dit wordt onder andere bevestigd door studies in Duitsland (Robitzsch et al, 2017) en voor wiskunde in Nederland (Feskens et al, in press). Beide studies tonen voor de desbetreffende landen via een alternatieve manier van equivaleren aan, dat de daling in landengemiddelden moderater uitvalt dan origineel door de OECD gerapporteerd.

Conclusie

Plausible values zijn zeer geschikt voor onderwijskundige surveys en leveren accurate populatieschattingen. Een cruciale stap bij het gebruik van meerdere toetsversies is dat de item parameters die nodig zijn om de moeilijkheid van verschillende toetsen in kaart te brengen, correct worden geschat. De overstap van een papieren naar een digitale afname in PISA-2015 heeft de validiteit van de plausible values onder druk gezet. De daling in scores in 2015 moet dan ook met meer dan de gebruikelijke voorzichtigheid geïnterpreteerd worden.

23 https://www.tes.com/news/exclusive-pisa-data-may-be-incomparable-schleicher-admits

Vaardigheids-ontwikkelingen volgens PISA en examens

Cito
Amsterdamseweg 13
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
klantenservice@cito.nl
www.cito.nl

Fotograaf: Gijs Versteeg