Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2012

Jaarlijks Peilingsonderzoek van het Onderwijsniveau

Jan van Weerden, Bas Hemker, Hendrik Straat en Kees Mulder

zeker weten

Jaarlijks Peilingsonderzoek van het Onderwijsniveau

Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2012

Jaarlijks Peilingsonderzoek naar het Onderwijsniveau
Jan van Weerden, Bas Hemker, Hendrik Straat en Kees Mulder

Cito
Arnhem, augustus 2013

Inhoud

Voorwoord

Dit rapport vormt de neerslag van de vijfde meting van de taal- en rekenvaardigheden van leerlingen in het basisonderwijs sinds 2008. De meting is van start gegaan naar aanleiding van de kwaliteitsagenda voor het Primair Onderwijs, verschenen in 2007 onder de titel ‘Scholen voor morgen’ en wordt sindsdien gecontinueerd als landelijke monitor voor een aantal basisvaardigheden.

Bij de uitvoering van dit onderzoek wordt gebruikgemaakt van data uit toetssystemen die al bij de meeste scholen worden gebruikt voor het in kaart brengen van onderwijsopbrengsten bij rekenen en taal. Het gaat dan om de Eindtoets in jaargroep 8 en toetsen van het Cito Volgsysteem voor primair onderwijs in jaargroep 4. Door deze keuze wordt de extra toetslast voor leerlingen en leerkrachten beperkt.

De uitkomsten van deze jaarlijkse peiling geven een indicatie voor de stand van zaken met betrekking tot een belangrijk deel van de referentieniveaus behorende bij de doorlopende leerlijnen zoals geformuleerd in het advies van de Expertgroep Doorlopende leerlijnen “Over de drempels met taal en rekenen”.

De resultaten voor jaargroep 8 worden meegenomen in het jaarlijkse verslag ‘De staat van het onderwijs’ van de Inspectie van het onderwijs.

Dit rapport zou niet mogelijk geweest zijn zonder de medewerking van de leerkrachten en schoolleiders van de basisscholen die bereid waren om hun gegevens ter beschikking te stellen van ons onderzoek. Wij danken hen hartelijk voor hun bijdrage aan het welslagen van deze peiling.

Bij de uitvoering van het project zijn diverse mensen betrokken geweest, waarvan er enkele met name genoemd moeten worden. De verspreiding van toetsmaterialen en de dataverzameling voor jaargroep 4 werd met zorg uitgevoerd door Gijs Marissink. Onder supervisie van Bas Hemker werden de telkens toch weer complexe analyses voor jaargroep 8 en jaargroep 4 uitgevoerd door respectievelijk Hendrik Straat en Kees Mulder.

We hopen dat dit rapport en de daarin beschreven resultaten hun weg vinden naar het onderwijsveld en de beleidsmakers.

Drs. J.J van Weerden
Projectleider PPON

Samenvatting

In dit rapport wordt verslag gedaan van de vijfde jaarlijkse peiling van de reken- en taalvaardigheid in de jaargroepen 4 en 8. Daarvoor is in jaargroep 8 gebruikgemaakt van gegevens uit de Eindtoets van 2012 betreffende Woordenschat, Spelling, Begrijpend lezen en Rekenen-Wiskunde. Voor jaargroep 4 is een afzonderlijk steekproef van scholen getrokken en zijn scholen voorzien van LVS-toetsen uit het Cito Volgsysteem primair onderwijs voor Woordenschat, Spelling, Begrijpend lezen en Rekenen-Wiskunde.

Vergelijken we de prestaties van 2012 met die van 2011, dan blijken die van de leerlingen in jaargroep 8 overwegend licht te zijn gestegen. Dat is bij Rekenen-Wiskunde het geval en bij twee van de drie taalvaardigheden, nl. bij Spelling en in mindere mate bij Woordenschat. Bij Begrijpend lezen is het beeld minder eenduidig voor wat betreft de vergelijking van 2012 met 2011. De hele periode van vijf jaar overziend blijkt dat er bij alle onderdelen – ook Begrijpend lezen – ten opzichte van de start van de meting in 2008 vooruitgang is te melden.
Bij jaargroep 4 zien we een vergelijkbaar beeld als bij jaargroep 8. De positieve effecten zien we vooral bij Rekenen-Wiskunde, Spelling en Woordenschat. Begrijpend lezen is ook hier een twijfelgeval: de prestaties lijken weer te dalen, maar blijven nog wel duidelijk boven het startniveau van 2008.

Nemen we de twee jaargroepen bij elkaar dan zien we dat er voor Rekenen-Wiskunde sprake is van een eenduidige positieve trend. Die is in jaargroep 4 groter dan in jaargroep 8, maar in beide gevallen onmiskenbaar.
Voor Spelling is eveneens in beide jaargroepen een eenduidige positieve ontwikkeling te melden die ook in jaargroep 4 wat sterker is dan in jaargroep 8. Woordenschat laat in jaargroep 4 eveneens een duidelijke stijging zien met dezelfde omvang als Spelling en Rekenen-Wiskunde, maar heeft in jaargroep 8 een grillig verloop, met uiteindelijk nauwelijks vooruitgang ten opzichte van 2008. Voor begrijpend lezen is de situatie nog lastiger te duiden. In jaargroep 4 is er nauwelijks progressie te zien vergeleken met 2008. Vanaf 2011 is er weer een lichte daling geconstateerd. In jaargroep 8 lijkt het resultaat in 2011 een uitschieter naar boven te zijn. Ten opzichte van dat jaar lijkt 2012 gedaald te zijn, maar het resultaat is vergelijkbaar met dat van 2010. Dat resultaat is nog steeds beter dan in 2008.

Alles bij elkaar genomen mogen we dus concluderen dat de trend in alle gevallen positief uitvalt, waarbij we de grootse vooruitgang aantreffen in jaargroep 4. Alleen de vaardigheid in Begrijpend lezen blijft in de buurt van het niveau van 2008. Een verklaring voor het verschil met de andere vaardigheden is nog niet gevonden.

Uit de gegevens blijkt dat jongens in jaargroep 8 beter presteren bij Rekenen-Wiskunde dan meisjes. Meisjes scoren hoger bij Spelling. Ook presteren zij significant beter bij Begrijpend lezen, maar daar is het verschil wel een stuk kleiner. In jaargroep 4 is dit verschil tussen jongens en meisjes ook significant en groter dan in jaargroep 8. Bij Woordenschat, waar jongens en meisjes in jaargroep 8 in 2012 gelijk presteerden, zien we dat in jaargroep 4 meisjes een hogere score behalen dan jongens. De verschillen bij Rekenen-Wiskunde en Spelling zijn vergelijkbaar met wat gevonden is in jaargroep 8. Het verschil tussen jongens en meisjes lijkt vrij constant, zeker in jaargroep 4. Opvallend is wel dat de verschillen in omvang in termen van vaardigheidsscores groter blijken te zijn in jaargroep 4 dan in jaargroep 8. Het lijkt erop dat het verschil ten gevolge van geslacht in de loop van de basisschool, in ieder geval van jaargroep 4 naar jaargroep 8, dus afneemt.

Er is geen onderzoek gedaan naar mogelijke oorzaken voor de uitkomsten over de jaren heen. Zo is, anders dan bij regulier peilingsonderzoek (PPON), geen zicht op het onderwijsaanbod of veranderingen daarin, laat staan dat dergelijke gegevens direct gekoppeld kunnen worden aan de leerlingprestaties.
Dat maakt het lastig om aan deze uitkomsten een interpretatie te verbinden. We weten dus niet op welke wijze bijvoorbeeld het gebruik van andere lesmethoden of het besteden van meer tijd aan bepaalde onderwerpen te maken heeft met deze resultaten. Om een relatie te leggen tussen de gesignaleerde vooruitgang en wijzigingen in onderwijsbeleid is nader onderzoek nodig naar tussenliggende factoren.

Inleiding

In het kader van de kwaliteitsagenda ‘Scholen voor morgen’ is in 2008 het Jaarlijks Peilingsonderzoek naar het Onderwijsniveau (JPON) van start gegaan. Dit jaarlijkse onderzoek is gericht op het monitoren van het onderwijsniveau op het gebied van taal- en rekenvaardigheid in het basisonderwijs. In 2008 is verslag gedaan van de eerste jaarlijkse niveaupeiling van taal en rekenen in jaargroep 8 en 4 (Hemker & Van Weerden, 2009). Het project is ondergebracht bij PPON, de periodieke peiling van het onderwijsniveau, maar kent een andere methodiek en gebruikt andere instrumenten. De overeenkomst is dat het ook hier gaat om periodiek onderzoek, zij het met een hogere frequentie, en dat ook hier resultaten op een vaste meetschaal worden gerapporteerd, die maakt dat de uitkomsten over de jaren heen vergelijkbaar zijn.
De belangrijkste verschillen zijn te vinden in de gebruikte instrumenten, de wijze van afnemen en de beschikbare achtergrondgegevens.

De kwaliteitsagenda ‘Scholen voor morgen’ was gericht op duurzame verbetering van het primair onderwijs en legde de prioriteit op de verhoging van de taal/lees- en rekenopbrengsten. Dit is gecontinueerd in het beleid van daaropvolgende kabinetten en heeft onder andere geleid tot nieuwe wetgeving en voorstellen daartoe, waarin de positie van rekenen en taal en het toezicht daarop verder is verstevigd.
In de aanbevelingen van de Expertgroep Doorlopende Leerlijnen (EGDLL, ofwel de commissie Meijerink, SLO, 2008) werd vastgelegd wat leerlingen moeten kennen en kunnen op dit gebied bij het verlaten van de basisschool. Dit staat bekend als de referentieniveaus voor taal en rekenen. Om vast te stellen wat de beginsituatie was en hoe het niveau zich sindsdien heeft ontwikkeld in het basisonderwijs is het Jaarlijks Peilingsonderzoek (JPON) gestart.

Scholen beschikken doorgaans over voldoende toetsgegevens, gebaseerd op leerlingvolgsystemen en eind- of entreetoetsen, als input voor verbeterplannen op groeps-, school- en bestuursniveau. Deze gegevens kunnen echter ook, onder bepaalde condities, worden ingezet voor een landelijke niveaubepaling, zoals dat in het project JPON wordt gedaan. Wel moet worden vastgesteld dat de bestaande toetssystemen inhoudelijk nog niet zijn afgestemd of ingericht op de referentieniveaus van de commissie Meijerink. De toetsen die nu ingezet kunnen worden zijn nog niet dekkend voor het hele terrein dat met de referentieniveaus wordt beschreven. Met name de productieve vaardigheden bij taal (schrijven en spreken) ontbreken nog, net als taalvaardigheden als begrijpend luisteren of grammaticaliteit (of kunnen omgaan met elementen uit de ‘begrippenlijst’). De toetsresultaten geven echter wel een duidelijke indicatie van de stand van zaken op een aantal belangrijke vaardigheden en maken het ook mogelijk daarin de ontwikkeling van jaar tot jaar te laten zien.

In 2008 is deze jaarlijkse peiling voor de eerste keer uitgevoerd. Daarmee is een start gemaakt met wat een reeks van jaarlijkse rapporten moet worden, waarin het niveau van taal en rekenen in het primair onderwijs in beeld wordt gebracht. Daarbij is 2008 steeds het uitgangspunt voor de vergelijking, behalve bij de toets Woordenschat in jaargroep 4. Deze toets werd pas in 2009 voor de eerste keer ingezet.

Het verslag begint met een beschrijving van de opzet van het onderzoek, de gebruikte toetsen en de wijze waarop deze toetsen met elkaar vergeleken kunnen worden. Vervolgens wordt de onderzoekspopulatie beschreven en worden de definities van de achtergrondvariabelen gegeven (hoofdstuk 1). Ook wordt beschreven hoe we de resultaten van de onderzoekspopulatie op de gemeten vaardigheden kunnen vergelijken met voorgaande jaren.
Vervolgens beschrijven we de resultaten van de leerlingen in jaargroep 8 (hoofdstuk 2). Daarna volgen de resultaten van jaargroep 4 (hoofdstuk 3). In beide hoofdstukken geven we een vergelijking over de jaren heen op de onderwerpen voor rekenen en taal, en de relatie van deze uitkomsten met enkele achtergrondvariabelen zoals geslacht en thuistaal.
Het laatste hoofdstuk bevat een samenvatting, conclusie en discussie.

Tot slot zijn er enkele bijlagen opgenomen, waarin voor elk onderwerp de uitkomsten worden gegeven voor 2012, plus bijbehorende effectgrootten en waarin verder ook de uitkomsten voor de diverse achtergrondvariabelen zijn getabelleerd. Deze tabellen kunnen naast die van de vorige peilingen, zoals opgenomen in eerdere verslagen, worden gelegd en gezien worden als aanvulling daarop.

1 Vraagstelling en aanpak

1.1 Vraagstelling

Het belangrijkste doel van JPON is het nauwkeurig vaststellen van veranderingen in de taal- en rekenvaardigheden van leerlingen in jaargroep 8 en 4 van het basisonderwijs. Daarnaast is het doel na te gaan in hoeverre de prestatieverschillen tussen bepaalde groepen leerlingen gelijk zijn gebleven dan wel groter of kleiner zijn geworden.

1.2 Methode

Voor deze peiling maken we gebruik van instrumenten die reeds in een ander kader worden ingezet. Het zijn dus geen nieuwe toetsen die afzonderlijk voor dit project zijn ontwikkeld. Voor jaargroep 8 maken we gebruik van de Eindtoets Basisonderwijs, een toets die een aantal onderdelen bevat voor rekenen en taal die we goed voor dit doel kunnen gebruiken, maar die elk jaar volledig wordt vernieuwd. Omdat deze toets elk jaar andere items bevat, kunnen de resultaten niet zomaar van jaar tot jaar met elkaar worden vergeleken. Door gebruik te maken van aanvullende databronnen is hier toch een goede oplossing voor gevonden. We beschrijven dit verderop in dit hoofdstuk (par. 1.2.1).

Voor jaargroep 4 zijn er geen data uit een landelijke toets beschikbaar. Wel wordt door veel scholen gebruikgemaakt van de LVS-toetsen van het Cito Volgsysteem voor primair onderwijs, het voormalige LOVS. Er is in dit geval een steekproef getrokken van scholen met de vraag of men de gegevens van een aantal LVS-toetsen aan ons wilde verstrekken. Om dat voor deze scholen mogelijk en ook aantrekkelijk te maken kregen ze de toetsmaterialen gratis toegezonden, met het verzoek de verkregen scores aan ons te retourneren. De LVS-toetsen worden echter niet elk jaar vernieuwd. Een voordeel hiervan is dat de meetschalen bekend zijn en niet meer veranderen. Doordat de inhoud van de toetsen onveranderd is, is het niet mogelijk dat wijzigingen in de inhoud van de toets een verklaring zijn voor de veranderingen over de tijd. Een nadeel is echter dat de bekendheid van het materiaal mogelijk van invloed kan zijn op de gevonden resultaten
De hier genoemde toetsen zijn in aantal items en mogelijk te onderscheiden eenheden lang niet zo gedifferentieerd als in het reguliere onderzoek van PPON gebruikelijk is. Het aantal rapportage-eenheden is daardoor beperkt tot een zevental onderwerpen in jaargroep 8 en een viertal onderwerpen in jaargroep 4. Ook worden bij PPON meer en andere itemvormen ingezet dan alleen de vierkeuzevragen van de Eindtoets of de meerkeuzevragen en kort-antwoordmodellen van de LVS-toetsen.

Een belangrijk verschil met PPON is verder de afnameconditie. Bij PPON worden alle toetsen en taken afgenomen door een getrainde toetsleider. In dit geval was de eigen leerkracht de toetsleider. Ook dat vormt een mogelijke aantasting van de objectiviteit van de afname.
Tot slot moet gewezen worden op het verschil in impact van de verschillende toetssituaties. De Eindtoets kan worden beschouwd als een ‘high-stake’ toets, een toets waar leerlingen hun uiterste best op zullen doen, niet alleen vanwege de gevolgen van de uitslag voor hun verdere toekomst, maar ook vanwege de grote mate van aandacht en publiciteit er om heen. Voor LVS-toetsen geldt dat in mindere mate en voor afnames voor PPON lijkt dat nog minder het geval. Deze verschillen in conditie en daardoor in motivatie van de leerlingen hebben hun weerslag op de hoogte van de toetsscores, zo is inmiddels uitgezocht (zie verder Hemker, 2013).

Voor een vergelijking over de jaren heen zijn deze risico’s van minder groot belang, omdat de condities immers in de meeste gevallen niet over de jaren heen zullen variëren. Het is dus geen probleem om een trend in beeld te brengen. Voor een vergelijking van deze uitkomsten met die van andere onderzoeken zijn ze dat echter wel. Zo zullen resultaten op de Eindtoets een ander beeld kunnen geven van de mate waarin de landelijk vastgestelde referentieniveaus zijn gehaald dan de uitkomsten van een LVS-meting of een PPON-meting.

Uit overwegingen van efficiëntie en kosten zijn voor de jaarlijkse niveaupeilingen dus overwegend gegevens gebruikt die toch al door scholen worden verzameld. Voor de meting van de taal- en rekenvaardigheid in jaargroep 8 zijn dat de verzamelde gegevens uit campagne van de Eindtoets Basisonderwijs 2012. Daarnaast is gebruikgemaakt van gegevens verzameld met de bijbehorende Niveautoets, de ankertoets en de nu voor de tweede keer op een aantal scholen afgenomen toets Basisvaardigheden. Voor taal betreft het de onderdelen Begrijpend lezen (BL), Spelling (Sp) en Woordenschat (Wo). Bij rekenen gaat het om de onderdelen Getallen en bewerkingen (GB), Breuken, procenten en verhoudingen (BPV) en Meten, meetkunde, tijd en geld (MMTG).

In jaargroep 4 is gebruikgemaakt van toetsen van het Cito Volgsysteem primair onderwijs (LOVS). Voor taal is gekozen voor dezelfde onderdelen als in jaargroep 8. Bij rekenen zijn in eerdere peilingen vier onderdelen onderscheiden in plaats van drie, namelijk Getallen en getalsrelaties (G/G), Optellen en aftrekken (O/A), Vermenigvuldigen en delen (V/D) en Meten, tijd en geld (MTG). Vanaf de peiling 2011 worden deze echter ook als samengestelde rekenschaal onderzocht. Dat is mogelijk aangezien de verschillende subvaardigheden (zeer) hoog met elkaar correleren. Bij de peiling van 2012 is deze samengestelde rekenschaal de enige schaal waarop de resultaten worden gegeven. Een steekproef van scholen is gevraagd om gegevens te leveren over deze toetsen voor jaargroep 4. Deze toetsen worden doorgaans afgenomen aan het eind van het schooljaar in de maanden mei-juni.
In beide gevallen, dus zowel in jaargroep 4 als in jaargroep 8, is sprake van een steekproefprocedure, zodat de resultaten als representatief mogen worden gezien voor het niveau in Nederland op de beide meetmomenten.

1.2.1 Kwaliteit van de meetinstrumenten

Er is gebruikgemaakt van de resultaten op de Eindtoets Basisonderwijs (EB) van 2012. Deze toets wordt elk jaar opnieuw samengesteld en bestaat uit opgaven die al in proeftoetsen in voorgaande jaren zijn uitgeprobeerd. De psychometrische kenmerken zijn van te voren goed in te schatten en blijken telkens van hoog niveau te zijn. De eigenschappen van de Eindtoets worden elk jaar gepubliceerd in een afzonderlijke publicatie en dat is ook voor 2012 het geval (Cito, 2012). We mogen hier stellen dat de psychometrische kwaliteit bij de Eindtoets gewaarborgd is door de gehanteerde procedures. Nadere informatie hierover is te vinden in de uitgebreide verantwoording van de Eindtoets van 2010 (Van Boxtel e.a., 2012).

We moeten wel wijzen op het feit dat de wijze waarop deze toets en de versies daarvan in dit onderzoek worden ingezet wel een andere is dan waarvoor zij is bedoeld. Het primaire doel van de Eindtoets is namelijk het opleveren van een goede prognose voor de kans van slagen in verschillende vormen van voortgezet onderwijs voor individuele leerlingen. In deze studie gaat het niet om de totaalscore voor alle onderdelen, de standaardscore, maar om de uitkomsten op delen van de toets. In JPON worden verschillende onderwerpen van de toets afzonderlijk gebruikt om een beeld te schetsen van de kwaliteit van het onderwijs op systeemniveau. Voor dat doel zijn de psychometrische eisen die aan een toets moeten worden gesteld anders. Met dat fenomeen is rekening gehouden in de analyses, op een vergelijkbare wijze als in voorgaande jaren (zie verder Hemker & Van Weerden, 2009; Hemker, Kuhlemeier & Van Weerden, 2010; Hemker, Kordes en Van Weerden, 2011).

Daar waar het de toetsen voor jaargroep 4 betreft kunnen we opmerken dat die niet veranderd zijn ten opzichte van vorige jaren en dat daarom ook de meeteigenschappen van de toetsen niet veranderd zijn. De toetsen voor jaargroep 4 maken deel uit van het Cito Volgsysteem primair onderwijs. Deze toetsen worden pas in productie genomen na een uitvoerige pretest en kunnen bogen op een goede psychometrische kwaliteit. De in dit onderzoek gehanteerde toetsen zijn regulier aangeboden aan de COTAN (Commissie Test Aangelegenheden Nederland) en zijn op de relevante criteria als voldoende en goed beoordeeld. De psychometrische kwaliteit is weergegeven in de respectievelijke wetenschappelijke verantwoordingen van deze toetsen (Van Berkel e.a., 2010; Feenstra e.a., 2010; Janssen e.a., 2010; De Wijs e.a., 2010). De eigenschappen van de toetsen in de context van JPON zijn in eerdere verslagen beschreven. Zie voor verdere informatie: www.toetswijzer.nl.

1.2.2 Steekproeftrekking en analyse

Jaargroep 8

In dit onderzoek zijn twee databestanden gebruikt om tot steekproeven te komen. Die steekproeven verschilden wel in grootte. De steekproef van jaargroep 8 is veel omvangrijker dan die van jaargroep 4. De steekproef in jaargroep 8 is gebaseerd op de data van het totaal aantal deelnemers aan de Eindtoets in 2012. Na een eerste schifting van leerlingen die buiten de onderzoekspopulatie vallen, resteerden 150.168 deelnemers. Achterwege zijn gelaten vooral leerlingen van scholen in het buitenland, speciaal onderwijs en voortgezet onderwijs en leerlingen die de digitale versie van de Eindtoets hadden gemaakt.
Vervolgens zijn uit het databestand twee steekproeven getrokken: één ten behoeve van de schaal-constructie en één voor de analyse met achtergrondvariabelen.
Voor de schaalanalyses (OPLM) is gebruikgemaakt van de resultaten van alle leerlingen die beschikbaar zijn in de koppeling tussen jaargangen. Bij het schalen van de opgaven is de representativiteit van de steekproef van minder groot belang, vanwege populatieonafhankelijkheid van de itemparameterschattingen bij de gebruikte analysemethode (item response theorie; IRT). In deze steekproef hebben we alle leerlingen ingesloten die de Niveautoets (NT) of de toets Basisvaardigheden (TBV) of de ankertoets hebben gemaakt. Vervolgens is die dataverzameling aangevuld met een aselecte steekproef uit alle overige leerlingen van de Eindtoets totdat er uiteindelijk 30.000 leerlingen in het bestand zaten. De NT-leerlingen werden ingesloten omdat zij een andere Eindtoets hebben gemaakt. De ankertoets en de TBV werden ingesloten om de toetsen op onderdelen te kunnen koppelen. Daarbij is het gebruik van data van de TBV in de analyse nieuw dit jaar. De kwaliteit van de gebruikte schalen is niet anders dan die in voorgaande jaren. Merk op dat met de gebruikte aantallen iedere kleine afwijking in veel gevallen als significant gekwalificeerd wordt maar dat deze afwijking geen effect heeft op de resultaten (zie voor meer hierover in de vorige JPON-rapportages). Na de schaalanalyse is met de geconstrueerde schalen, die in principe dezelfde psychometrische kenmerken bezaten als in voorgaande jaren, de analyse gedaan op achtergrondvariabelen, met in de eerste plaats de jaarvergelijking. Daartoe is een nieuwe steekproefprocedure uitgevoerd waarbij 5 aselecte steekproeven zijn getrokken van ongeveer 30.000 leerlingen zonder teruglegging, waarbij representativiteit wel van belang is. De resultaten in dit rapport zijn gebaseerd op de meest representatieve steekproef van de vijf. Om de robuustheid van deze steekproef te bepalen zijn de SAUL-analyses op drie verschillende steekproeven uitgevoerd. Er bleek weinig steekproeffluctuatie te zijn. Uiteindelijk is voor de verdere analyses gerekend met de gegevens van 29.987 leerlingen. De kenmerken van deze groep op achtergrondvariabelen zijn volkomen vergelijkbaar met de totale groep en die zijn terug te vinden in de terugblik bij de Eindtoets (Cito, 2012). We mogen dus stellen dat deze steekproef een representatief beeld geeft van de Nederlandse populatie in jaargroep 8.

Jaargroep 4

Voor jaargroep 4 is een afzonderlijke steekproef getrokken uit het scholenbestand van Cito. Dit betrof een gestratificeerde steekproef met strata gebaseerd op de formatiegewichten van de scholen, conform de procedure bij PPON, waarbij werd gemikt op ca. 100 scholen. Scholen die al participeerden in andere onderzoeken van Cito werden buiten de steekproef gehouden. De scholen werden aangeschreven in februari 2012 en kregen bij deelname gratis toezending van de benodigde toetsmaterialen. Deze waren ook bruikbaar en beschikbaar voor administratie in de rapportagemodule van het Cito Volgsysteem. Als tegen-prestatie leverden de scholen een ingevuld digitaal scoringsblad in na afloop van de toetsafname in de maanden mei-juni. De werving leverde uiteindelijk 87 scholen op met in totaal 2394 leerlingen waarvan de achtergrondgegevens bekend waren.
Doordat de gebruikte toetsen onveranderd waren en in de eerdere peilingen de schaling van de opgaven al was uitgevoerd is er bij de data-verzameling gebruik gemaakt van de totaalscores op de verschillende toetsonderdelen. Ook waren de toets- en itemeigenschappen van de toetsen al bekend. De itemscores zijn zodoende niet meer opgevraagd, maar zijn rechtsreeks uit de totaalscores verkregen. Deze zijn door de leerkracht opgestuurd, samen met informatie over welke (versie) van de toetsen de leerlingen gemaakt hadden en de achtergrondgegevens van de leerlingen. Op basis van de informatie uit de eerder uitgevoerde peilingen kunnen de scores van de verschillende toetsen op de vaardigheidsschalen geplaatst worden. Bij de toetsen Begrijpend lezen en Spelling wordt met aparte – in moeilijkheid van elkaar verschillende – versies gewerkt. Bij deze toetsen moet naast de somscore ook bekend zijn welke versie de

leerling gemaakt heeft voordat een latente trek kan worden toegekend op basis van de somscore. Met dit fenomeen is rekening gehouden.

Zoals in de beschrijving van de methode in paragraaf 1.2 al is aangegeven, is de analyse voor 2012 anders uitgevoerd dan in vorige jaren. Er is nu direct gewerkt met de totaalscores op de toets, zonder dat de gegevens op itemniveau beschikbaar waren. Deze totaalscores zijn door de leerkrachten ingevuld op een scoreblad, samen met de variabelen voor de achtergrondkenmerken en de gegevens over de versie die leerlingen gemaakt hebben.

Bij de rekenschalen is nu gebruikgemaakt van de samengestelde rekenschaal. Dit kan zonder bezwaar, aangezien de verschillende mogelijke subschalen voor rekenen zeer hoog met elkaar correleerden (correlatie was gemiddeld hoger dan 0,90). De resultaten op de verschillende subschalen lijken daardoor ook zeer sterk op elkaar en konden zodoende goed samengevat worden in de samengestelde rekenvaardigheidsschaal. De correlaties voor de verschillen schalen bij taal waren aanzienlijk lager (gemiddeld 0,66) waardoor informatie verloren zou gaan als alleen die taalschaal gerapporteerd zou worden. De resultaten op een algemene schaal voor taal zou ook erg afhankelijk zijn van de specifieke samenstelling van deze schaal.
De analyses van de effecten van de achtergrondvariabelen zijn gedaan op de vaardigheidsscores, zodat de gerapporteerde resultaten vergelijkbaar zijn met die van de eerdere rapportages. Naast de geschatte gemiddelden (en andere verdelingseigenschappen) van latente variabelen is er bij de analyse gebruikgemaakt van een “Generalized Lineair Model” (GLM) met hoofdeffecten. Met deze methode kunnen, vergelijkbaar met de eerdere uitgevoerde analyses met SAUL, de gecorrigeerde effecten berekend worden.

Zowel voor jaargroep 8 als jaargroep 4 zijn de steekproeven een goede afspiegeling van de populaties van leerlingen.

1.2.3. Achtergrondvariabelen

Behalve de genoemde taal- en rekenvaardigheden zijn ook een aantal achtergrondkenmerken in het onderzoek betrokken. Deze zijn te onderscheiden in leerlingkenmerken en schoolkenmerken.

Leerlingkenmerken:
- geslacht: jongens-meisjes;
- leertijd: leerlingen die al of niet eens hebben gedoubleerd, resp. regulier en vertraagd;
- formatiegewicht: gewicht van de leerling voor de formatieregeling op grond van opleiding en herkomst (alleen bij de oude regeling) van de ouders; verdeeld in drie categorieën, geen gewicht (0.00), laag gewicht (0.30) en hoog gewicht (1.20);
- thuistaal: Nederlands gesproken, een andere taal of een combinatie;
- advies VO: ingeschat niveau vervolgonderwijs door leerkracht (alleen jaargroep 8);
- wel of niet een IJK-code (alleen jaargroep 8), wat staat voor respectievelijk: (Allochtone) leerlingen die aan het begin van jaargroep 8 vier jaar of korter in Nederland zijn en die het Nederlands onvoldoende beheersen om de opgaven in de Eindtoets goed te kunnen lezen (Code I), of leerlingen die naar verwachting naar het (voortgezet) speciaal onderwijs of naar het praktijkonderwijs (pro) gaan (Code J) of naar verwachting in aanmerking komen voor het leerwegondersteunend onderwijs (lwoo) (Code K).
- gemaakte toets: Eindtoets of Niveautoets.

Schoolkenmerken:
- stratum: schoolindeling op basis van het percentage leerlingen met een formatiegewicht, verdeeld in drie categorieën; 1= weinig (<10%), 2= matig (10-25%) en 3 = veel (>25%);
- regio (noord, oost, west, zuid);
- urbanisatiegraad van de locatie van de school;
- schoolgrootte (alleen jaargroep 8).

De achtergrondvariabelen zijn niet in beide gepeilde jaargroepen hetzelfde. Voor jaargroep 4 was uiteraard het advies VO niet beschikbaar, evenmin als de IJK-codering, en was ook de schoolgrootte niet bekend. Ook zijn in de loop van de tijd variabelen veranderd, zoals de definitie van het formatiegewicht en van

stratum. Hoe hier mee om is gegaan is in eerdere verslagen van de peilingen beschreven. Dit jaar is de variabele IJK voor het eerst meegenomen in de analyses bij gebrek aan de variabele thuistaal die niet opgevraagd was de in de Eindtoets-campagne van 2012.

De achtergrondvariabelen worden bij de Eindtoets verzameld via het antwoordblad dat bij deze toets hoort en waarop leerlingen en leerkrachten een aantal gegevens invoeren. Bij de steekproef in jaargroep 4 werden de achtergrondgegevens per leerling op een afzonderlijke leerlingenlijst ingevuld door de betrokken leerkracht.

Opgemerkt moet worden dat het feit dat de wijziging in definitie van formatiegewichten in de loop van de in dit onderzoek bestreken periode ook een rol speelt bij de vergelijking over de jaren. Tevens heeft deze wijziging gevolgen gehad voor de indeling in de strata van de scholen. Hoe daarmee is omgegaan is uitvoerig toegelicht in de rapportage van het peilingsjaar 2010 (Hemker, Kordes & Van Weerden, 2011).

1.3 Hoe is er geanalyseerd?

Hieronder geven we beknopt weer hoe we de resultaten van 2012 hebben vergeleken met die van 2011 en de jaren daarvoor. Voor een uitgebreide technische verantwoording van de steekproef, de gebruikte toetsen, de statistische analyse en de rapportagemethodiek wordt verwezen naar de technische rapportages van eerdere jaren (Hemker & Van Weerden, 2009; Hemker, Kuhlemeier en Van Weerden, 2010; Hemker, Kordes & Van Weerden, 2011). Wel gaan we in op de afwijkingen in methodiek die er dit peilingsjaar zijn toegepast.

1.3.1 Vergelijking tussen en over de jaren

Een voorwaarde om vaardigheden van de verschillende jaren direct met elkaar te kunnen vergelijken, is dat de vaardigheden op dezelfde schaal gemeten zijn. Aan deze voorwaarde is voldaan als de leerlingen in de steekproef van 2012 precies dezelfde toetsen hebben gemaakt als de leerlingen in eerdere steekproeven. Voor jaargroep 4 is aan deze voorwaarde voldaan, aangezien de leerlingen in beide gevallen ongewijzigde toetsen uit het reguliere Cito Volgsysteem hebben gemaakt. De voor jaargroep 8 gebruikte Eindtoets Basisonderwijs wordt echter ieder jaar volledig vernieuwd. Wij hebben de vergelijkbaarheid echter kunnen waarborgen doordat de achtstegroepers telkens een aantal extra opgaven zijn voorgelegd (namelijk via ankertoetsen), die wel over de jaren heen ongewijzigd zijn gebleven. Ter aanvulling op die gegevens kon ook gebruik worden gemaakt van een overlap in items met de Niveautoets en de toets Basisvaardigheden. Deze laatste toets werd voor het eerst in 2011 op een groot aantal scholen in jaargroep 8 afgenomen en is in 2012 opnieuw gebruikt.

Met behulp van een speciale analysetechniek: het One Parameter Logistic Model (OPLM) (Verhelst, 1993; Verhelst en Glas, 1995), een variant van het item response model, zijn schalen geconstrueerd voor alle afzonderlijke onderwerpen. Dit is reeds in 2008 en 2009 uitgevoerd, waarbij de toetsen allen op dezelfde schaal werden gezet, met een gemiddelde van 250 punten en een standaardafwijking van 50. Op deze wijze kunnen de prestaties op de verschillende toetsen over de jaren heen met elkaar worden vergeleken (vgl. Hemker, Kordes & Van Weerden, 2011).

Voor het vergelijken van prestaties over de jaren heen moet ook rekening gehouden worden met de samenstelling van de responsgroep. Als de samenstelling gelijk is dan is een directe vergelijking mogelijk, maar in het geval van verschillen in samenstelling moet onderzocht worden waar die wijziging vandaan zou kunnen komen. Veranderingen in de samenstelling van een responsgroep kunnen het gevolg zijn van zogenaamde steekproeffluctuaties en van ‘echte’ veranderingen in de samenstelling van de populatie. Het probleem van steekproeffluctuaties zal zich voor jaargroep 8 niet zo gauw voordoen. Er is namelijk in elk jaar een zeer grote aselecte steekproef uit een bestand van vele honderdduizenden leerlingen getrokken (telkens ongeveer 85% van de populatie). Wel kan er zich een wijziging in de samenstelling van de populatie hebben voorgedaan. Stel dat alle basisscholen bijvoorbeeld minder leerlingen zijn gaan verwijzen naar het speciaal onderwijs. De responsgroep in 2012 zou dan meer ‘zorgleerlingen’ bevatten

dan die in 2011. We zouden dan ten onrechte kunnen concluderen dat de vaardigheid van de leerlingen achteruit is gegaan. Gelukkig zijn er statistische technieken beschikbaar die ons voor dit soort verkeerde conclusies kunnen behoeden. Vandaar dat wij voor jaargroep 8 zowel ongecorrigeerde als gecorrigeerde gegevens verstrekken (d.w.z. gecorrigeerd voor veranderingen in de samenstelling van de populatie). In dit geval wordt gebruikgemaakt van een speciale programma voor regressieanalyse van vaardigheidsscores geschat met behulp van OPLM, te weten SAUL: Structural Analysis of Univariate Latent variabels (Verhelst & Verstralen, 2002).

In jaargroep 4 is de steekproef veel minder groot dan in jaargroep 8 (namelijk 'slechts' ongeveer 2400 leerlingen van ongeveer 90 scholen). De omvang van de te vergelijken steekproeven in jaargroep 4 is te klein om steekproeffluctuaties met voldoende zekerheid te kunnen vaststellen en hiervoor vervolgens statistisch te kunnen corrigeren. Wel is het mogelijk te corrigeren voor veranderingen in de samenstelling van de populatie van vierdegroepers. Hierbij zijn zowel gecorrigeerde als ongecorrigeerde analyses uitgevoerd. De achtergrondkenmerken waarvoor gecorrigeerd wordt, zijn in beide modellen geslacht, leertijd, stratum en formatiegewicht.

1.3.2 Analyses voor de leerlingen in 2012

De gegevens van 2012 zijn als afzonderlijke groep geanalyseerd. In dit rapport zijn de gegevens van de gemiddelden van de verschillende te onderscheiden groepen vergeleken op de vaardigheidsschalen. Daarnaast zijn ook hier gecorrigeerde effecten geschat. Door deze gecorrigeerde effecten is het mogelijk een inschatting te maken van het additieve effect van een variabele, zoals het aanvullende effect van de gesproken thuistaal, naast het effect dat al gevonden wordt op basis van de vooropleiding van de ouders, zoals gerepresenteerd in het formatiegewicht.
De achtergrondkenmerken waarvoor in deze modellen gecorrigeerd wordt in een hoofdeffecten-model, zijn geslacht, leertijd, stratum en formatiegewicht. In aanvullende analyses zijn de overige variabelen, zoals regio en thuistaal toegevoegd (een extra variabele per analyse). Doordat de vaardigheden in 2012 op dezelfde schaal liggen als de voorgaande jaren zijn de resultaten hiermee direct vergelijkbaar over de jaren heen.

In de eerdere verslagen wordt dieper ingegaan op verschillen tussen de gecorrigeerde modellen en de niet gecorrigeerde modellen, en de verschillen in interpretatie. Ook de gevolgen voor de vergelijking over jaren heen en binnen het analysejaar zijn daar uiteengezet.

1.3.3 Significantie en effectgrootte

Of een gemiddeld vaardigheidsverschil tussen twee jaren statistische significantie oplevert, hangt in belangrijke mate af van de steekproefgrootte. Hoe groter de steekproef, hoe eerder een verschil statistisch significant is. Voor jaargroep 8 is de steekproef zeer veel groter dan voor jaargroep 4. Om de resultaten toch zinvol met elkaar te kunnen vergelijken rapporteren wij behalve de statistische significantie ook de zogeheten effectgrootte. De effectgrootte wordt in ons geval berekend als het verschil tussen de gemiddelden van de twee jaren (of twee subgroepen) gedeeld door de (gepoolde) standaardafwijking van de twee groepen die onderling worden vergeleken. Bij de interpretatie van de effectgrootte hanteren we de vuistregel van Cohen (1988) die is afgebeeld in tabel 1.1. Alles met een effectgrootte boven de 0.20 noemen we hier betekenisvol.

Tabel 1.1 Kwalificatie van effectgrootten

Effectgrootte (zowel plus als min)	**Kwalificatie**
0,0 tot 0,2	geen effect
0,2 tot 0,5	klein effect
0,5 tot 0,8	matig effect
0,8 of groter	groot effect

1.3.4 De gekozen rapportageschaal

Elke vaardigheid in dit onderzoek is getransformeerd naar een schaal met een gemiddelde van 250 en een standaarddeviatie van 50. Dit is conform de werkwijze bij PPON (zie bijv. Janssen, Van der Schoot & Hemker, 2005). De startwaarde is voor elke schaal het gemiddelde dat we in 2008 hebben aangetroffen. Dat gemiddelde is op 250 gesteld (zie verder Hemker & Van Weerden, 2009). De transformatie heeft als voordeel dat we de prestaties voor verschillende vaardigheden en voor verschillende jaren naast elkaar kunnen zetten op een en dezelfde schaal. Om de resultaten goed te kunnen beschrijven richten we ons in de rapportage niet alleen op de gemiddelde leerling, maar ook op andere groepen leerlingen in de vaardigheidsverdeling. Voor dit doel zijn vijf typische leerlingen gedefinieerd gekoppeld aan kenmerkende percentielpunten (zie tabel 1.2). In de rapportage gaan we uit van het basisjaar en rapporteren in de jaren daarop de eventuele wijziging in percentielwaarde. Zo zal bij een eventuele vooruitgang in vaardigheid blijken dat meer leerlingen de schaalwaarde 250 bereiken, maar ook kunnen meer leerlingen de schaalwaarde 186 bereiken. Teruggerekend betekent dat een wijziging in het percentage leerlingen dat tot een groep typische leerlingen behoort ten opzichte van 2008.

In dit verband moet worden opgemerkt dat aan de schaalwaarde horende bij de laag vaardige leerling een speciale betekenis kan worden toegekend. Dit is het punt op de vaardigheidsschaal dat door minstens 75% van de leerlingen wordt bereikt. In het advies van de Expertgroep Doorlopende Leerlijnen Rekenen en Taal (SLO, 2008) wordt dit punt regelmatig genoemd als referentieniveau voor 1F. Het referentieniveau 1S zou dan voor Rekenen overeenkomen met de P50. Bij Taal is daarvoor de P75 genoemd. Gekoppeld aan de specifieke inhoud van de schalen die we hier rapporteren zijn er echter nog geen uitspraken gedaan over het gewenste niveau.

Tabel 1.2 Definiëring typische leerlingen in de vaardigheidsverdeling

Aanduiding	**Afkorting**	**Percentiel 2008**	**Schaalwaarde**
Zeer laag vaardig	ZLV	P10	186
Laag vaardig	LV	P25	216
Gemiddelde/standaard	G/S	P50	250
Hoog vaardig	HV	P75	284
Zeer hoog vaardig	ZHV	P90	314

2 De resultaten voor jaargroep 8

2.1 De vergelijking over de jaren

Ervan uitgaande dat steekproeffluctuaties geen rol spelen (zie paragraaf 1.2), vergelijken we in deze paragraaf de resultaten van jaargroep 8 in 2012 met voorgaande jaren tot en met 2008. We presenteren hier dus de ongecorrigeerde verschillen.
In tabel 2.1 zijn de uitkomsten voor de vijf opeenvolgende jaren weergegeven. Het gemiddelde voor 2008 is per definitie 250 en de standaarddeviatie is 50. De schaalwaarde van 250 is vastgelegd in 2008 en vormt het startpunt waarmee we de uitkomsten van 2009 en later vergelijken (zie verder Hemker & Van Weerden, 2009). Ook is weergegeven welke percentielscore daarbij hoort, dat wil zeggen het percentage leerlingen met een gelijke of lagere score dan 250. Als het verschil tussen 2012 en 2010 statistisch significant is, is het effect vetgedrukt. Evenzo geldt dat voor het verschil tussen 2012 en 2008.

Bij Taal zien we dat jaargroep 8 in 2012 op alle drie variabelen beter presteert dan in 2008, maar bij Woordenschat is dat niet significant. Bij Begrijpend lezen zagen we tot 2012 een continu stijgende trend, maar in 2012 is die weer naar beneden gebogen. Dat blijkt uit de vergelijking van de twee jaren door middel van de resultaten op de ankertoets zoals gebruikelijk tot nu toe. Een alternatief zou zijn om de resultaten van de twee jaren met elkaar te vergelijken met behulp van de toets Basisvaardigheden. In dat geval zou er sprake zijn van een stijging. In de beschrijving van de resultaten bij Taal wordt verder ingegaan op de resultaten bij Begrijpend lezen.

Woordenschat bereikte in 2010 een hoge score, maar zakt in 2011 weer terug en is nog steeds nog maar net hoger dan in 2008. Spelling levert in 2012 een duidelijk hogere waarde op dan in 2011. We zien deze verschuivingen ook terug in de percentielscores in tabel 2.3.
Bij Rekenen zien we bij alle variabelen dat de stijging van 2011 doorzet in 2012. In dit geval is de vooruitgang van 2011 naar 2012 het grootst bij Breuken, procenten en verhoudingen. Getallen en bewerkingen is in verhouding het minst gestegen, maar komt in 2012 toch ook 5 punten hoger uit dan in 2008.

*Tabel 2.1 Effectgroottes en gemiddelde schaalwaarden jaargroep 8**

	2012-2011	2012-2008	2008	2009	2010	2011	2012
T: Woordenschat	0,02	0,05	250	249	257	251	252
T: Spelling	**0,07**	**0,10**	250	250	252	252	256
T: Begrijpend lezen (anker)	**-0,09**	**0,06**	250	252	254	257	253
T: Begrijpend lezen (toets basisv.)	**0,10**	**0,25**	250	252	254	257	262
RW: Rekenen: over-all	**0,06**	**0,12**	250	250	253	253	256
RW: Getallen en bewerkingen	**0,07**	**0,11**	250	250	252	252	255
RW: Breuken, procenten en verhoudingen	**0,07**	**0,14**	250	250	254	253	257
RW: Meten, tijd en geld	**0,05**	**0,12**	250	249	254	254	256

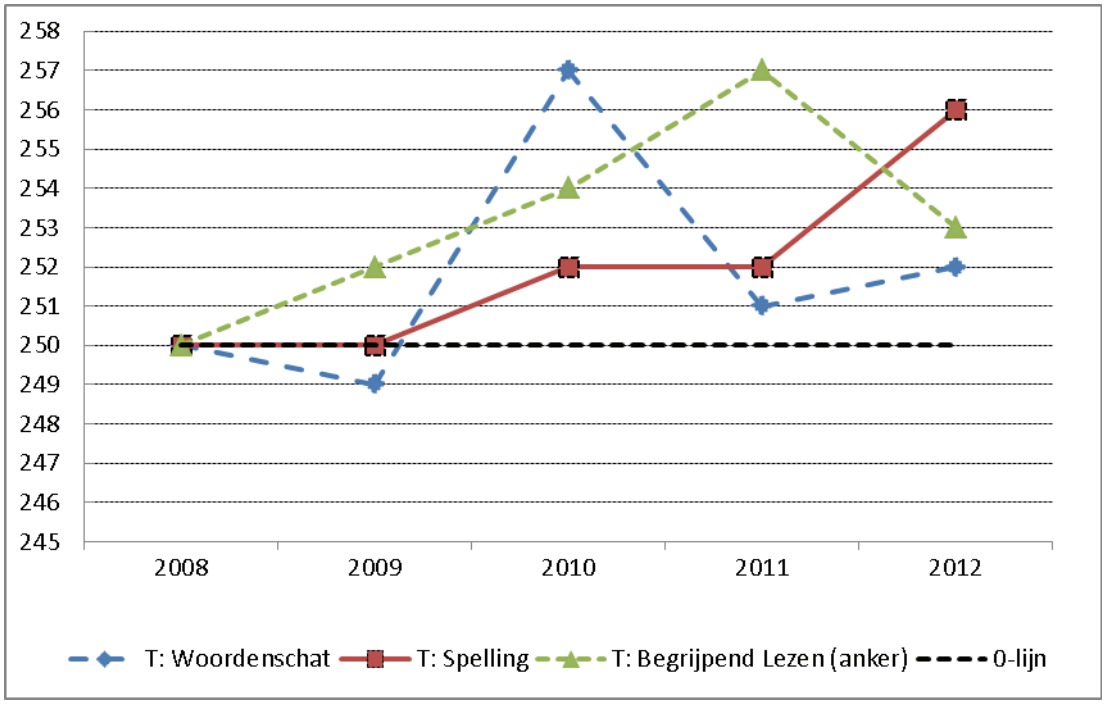
*Vet = significant verschil; **standaarddeviaties variëren van 49 tot 53; in 2008 per definitie 50.

2.1.1 Taal

Bij de taalonderwerpen zien we dat het behaalde niveau van de gemiddelde leerling in 2012, bijvoorbeeld bij Spelling, nu uitkomt op 256 op de meetschaal die is vastgezet op de waarden van 2008, een verschil van 6 punten. Ook bij Woordenschat zien we een verschil, maar dan gaat het om slechts 2 punten. Voor Begrijpend lezen is de situatie diffuus en zien we afhankelijk van de gekozen interpretatie een daling van 4 punten of een stijging van 5 punten. We komen daar dadelijk op terug. De verschillen over vijf jaren zijn in beeld gebracht met een trendlijn.

Opvallend is daarbij het grillige verloop van de scores op zowel Woordenschat als Begrijpend lezen (in de behoudende variant). Wel is duidelijk dat in alle gevallen 2012 een positief verschil laat zien ten opzichte van de beginsituatie van 2008.

Figuur 2.1 Trends over de jaren voor de taalonderwerpen in jaargroep 8



Het niet-lineaire verloop van de trend in de resultaten bij Woordenschat is al nader besproken in de rapportage van 2011, waarbij gewezen is op het verschil tussen jongens en meisjes bij de uitkomsten.

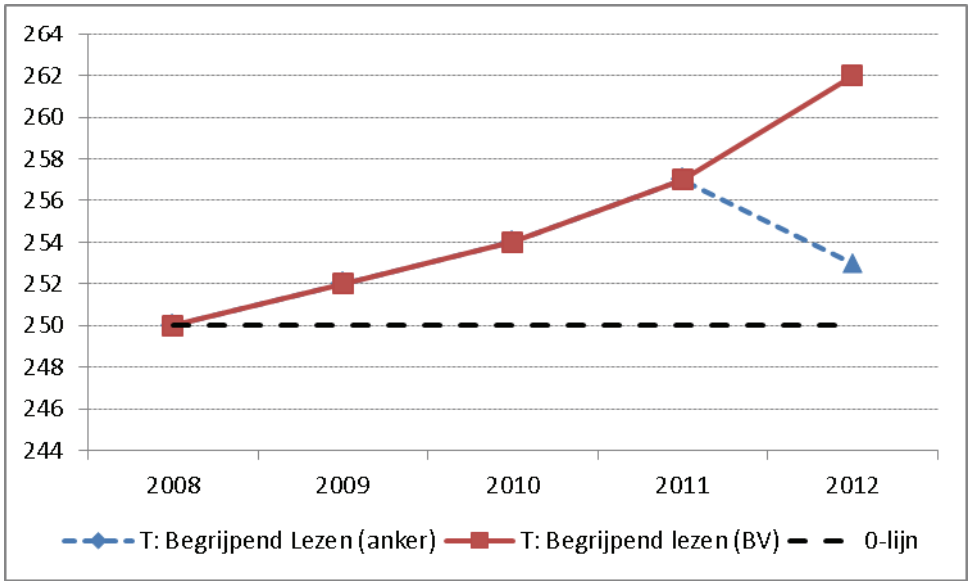
Bij de resultaten van Begrijpend lezen speelt een ander fenomeen. In figuur 2.1 worden nu de uitkomsten gebaseerd op een link via de ankertoets gerapporteerd. We zien zo een daling in 2012 na een gestage opwaartse trend van 2008 tot 2011. Dit lijkt voor een belangrijk deel verklaard te kunnen worden door een atypische, zij het wel mogelijke, uitkomst voor de ankergroep in 2011 die tot de (mogelijke) overschatting van de vaardigheid heeft geleid.

Indien we de analyse niet baseren op ankertoets, maar jaren met elkaar vergelijken door middel van de toets Basisvaardigheden, ontstaat er een ander, voor 2012 positiever beeld. Ter illustratie laten we in figuur 2.2 de trend zien voor beide benaderingen, dus zowel van de ankertoets (anker) als van de toets Basisvaardigheden (BV). De toets Basisvaardigheden maakt deel uit van het Cito Volgsysteem primair onderwijs, een toets die ook aan het einde van de basisschool wordt afgenomen, evenals de Eindtoets. Alleen zijn er twee afnamemomenten mogelijk: of nagenoeg tegelijkertijd met de Eindtoetsafname in de maanden januari-februari, of later in de maanden mei-juni. De toets Basisvaardigheden heeft een veel kortere traditie dan de Eindtoets en is in 2012 dan ook relatief nieuw. Het belang dat aan deze toets gehecht wordt lijkt echter door de jaren toe te nemen. Daar waar deze toets in 2010 vooral in een experimentele setting is gebruikt (die resultaten zijn toen ook niet voor JPON gebruikt) vonden de eerste meer serieuze afnamen plaats in 2011. In welke mate er door de leerlingen en scholen veel belang aan de afname gehecht wordt, is niet duidelijk. Vooralsnog lijkt het er op dat de toets in 2012 veel serieuzer gemaakt is dan in 2011. De gevonden stijging in resultaten bij Begrijpend lezen zou daardoor verklaard kunnen worden. Het betere resultaat is dan niet representatief voor een stijging in vaardigheid, maar de toename van het belang dat gehecht wordt aan de toets Basisvaardigheden (met andere woorden: de toets is meer high-stake geworden).

Vooralsnog wordt het meeste belang gehecht aan de resultaten zoals gevonden met de ankertoets. Dat is ook de wijze waarop de eerdere jaarvergelijkingen gedaan zijn. De vergelijking met de toets

Basisvaardigheden was eerder ook niet mogelijk. Een alternatieve stellingname is uit te gaan van het gemiddelde van de twee vergelijkingsmethoden, hetgeen een vrijwel ongewijzigd vaardigheidsniveau van Begrijpend lezen van 2011 naar 2012 oplevert. Een volgende meting in 2013 zal moeten uitwijzen welke benadering het meest verdedigbaar is.

Figuur 2.2 Trends over de jaren voor twee analyse van Begrijpend lezen in jaargroep 8



In tabel 2.2 hebben we ook de effectgroottes weergegeven voor elk contrast van jaar op jaar en voor de gehele periode. We zien daar dat het effect van Woordenschat van 2012 op 2011 niet significant is. In voorgaande jaren was dat wel het geval, maar in tegengestelde richting. Over de hele periode genomen is er geen significant verschil geconstateerd. Bij Spelling is zowel het verschil in het laatste jaar als voor de hele periode significant. Bij Begrijpend lezen is dus sprake van twee varianten. Daar is ook het effect van jaar op jaar telkens significant, maar in het contrast voor het laatste jaar is de richting echter verschillend, zowel positief als negatief.

In tabel 2.3 zien we de uitkomsten op de taaltoetsen op een andere wijze in kaart gebracht. Zo zien we dat de hogere schaalwaarde bij Begrijpend lezen (anker) zich bij de weergave bij de percentielen vertaalt in een percentage van 52 in 2012 ten opzichte van de standaardwaarde 50 van het jaar 2008. Dat betekent dat er nu 2 procent meer leerlingen een hoger waarde dan 250 hebben behaald. Als we de alternatieve analyse van Begrijpend lezen gebruiken, met de items uit de toets Basisvaardigheden, dan komen we echter op een percentage van 59 in plaats van 50, een stijging van 9%.

Tabel 2.2 Effectgroottes voor de jaarvergelijkingen taalonderwerpen jaargroep 8

	Woordenschat	Spelling	Begrijpend lezen (a)	Begrijpend lezen (b)
2009 - 2008	-0,02	0,00	**0,04****	-
2010 - 2009	**0,15****	**0,03***	**0,04****	-
2011 - 2010	**-0,11****	0,00	**0,07****	-
2012 - 2011	0.02	**0,07****	**-0,09****	**0,10****
2012 - 2008	0,05	**0,10****	**0,06****	**0,25****

Significantie: * a < ,01; ** a < ,001

Tabel 2.3 Vergelijking over de jaren heen in percentage per leerlingengroep voor de taalonderwerpen in jaargroep 8

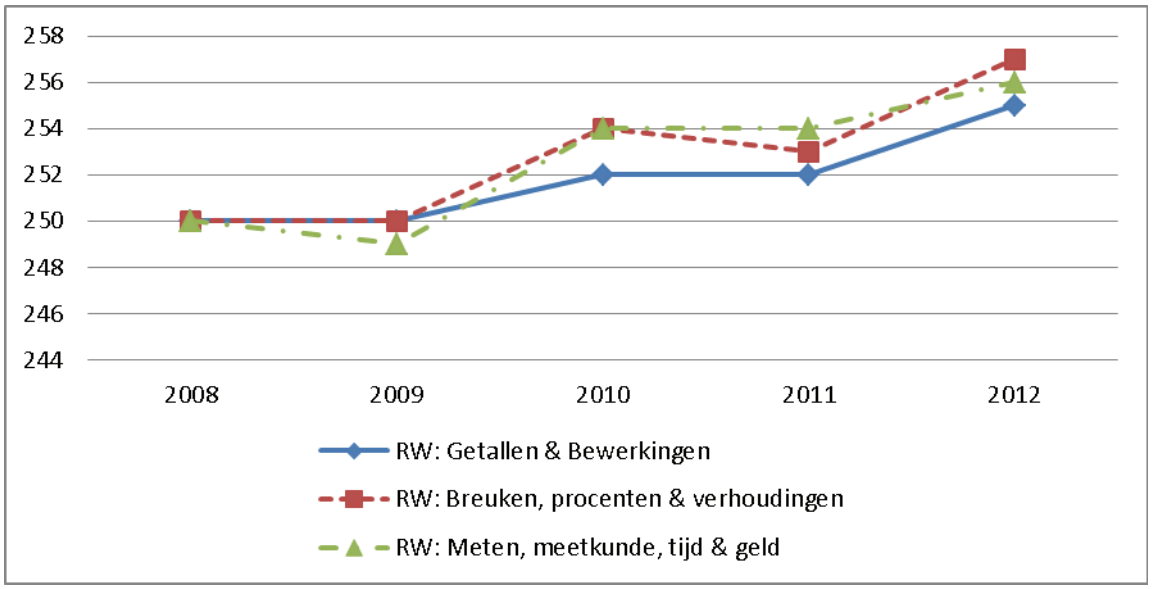
	Percentiel-waarde	Woordenschat				Spelling				Begrijpend lezen**				
Leerlingengroep*	**2008**	**09**	**10**	**11**	**12**	**09**	**10**	**11**	**12**	**09**	**10**	**11**	**12a**	**12b**
> dan ZLV	**10**	90	92	90	91	90	91	91	92	91	91	92	91	94
> LV	**25**	74	79	76	76	75	76	76	79	76	77	79	77	82
> Standaard 2008	**50**	49	55	51	52	50	51	51	55	52	53	56	52	59
> HV	**75**	24	29	26	26	25	26	26	29	26	27	30	27	33
>ZHV	**90**	10	13	10	11	10	11	11	12	11	11	13	11	15

* zie tabel 1.2; Begrijpend lezen **: 12a = anker; 12b = toets Basisvaardigheden

2.1.2 Rekenen-Wiskunde

Voor Rekenen-Wiskunde is het mogelijk om zowel een totaalschaal te presenteren als drie afzonderlijke schalen: voor elk onderscheiden onderwerp een. De onderlinge samenhang is dermate groot dat hier statistisch geen enkel bezwaar tegen is. We presenteren hier de afzonderlijke schalen. De sterke onderlinge samenhang van de drie onderwerpen zien we in de grafische weergave terug. Over de hele linie zien we hetzelfde beeld: in 2010 gaan alle schalen enkele punten omhoog, in 2011 is er nauwelijks verschil, maar in 2012 is er weer een duidelijke vooruitgang. Over de hele linie bekeken is de uitkomst in 2012 evident hoger dan in 2011 en dat is ook zo ten opzichte van 2008.

Figuur 2.3 Trends over de jaren voor de rekenonderwerpen in jaargroep 8



Bij de effectgroottes, weergegeven in tabel 2.4, zien we dat er een positief verschil is dat in alle drie gevallen significant is, maar de kwalificatie betekenisvol niet haalt. Als we de schalen samennemen is het contrast over de hele periode ook significant, maar is er een effectgrootte van 0,12. Dat ligt nog ver verwijderd van de 0,20, maar opvallend is wel dat de richting voor alle drie schalen positief is.
De percentielwaarden zijn met twee tot zes punten gestegen ten opzichte van 2008 zo blijkt uit tabel 2.5.

Tabel 2.4 Effectgroottes voor de jaarvergelijkingen rekenonderwerpen jaargroep 8

Contrasten	Onderwerpen			
	G/B	BPV	MMTG	Samen
2009 - 2008	-0,01	-0,01	-0,01	-0,01
2010 – 2009	**0,05****	**0,08****	**0,07****	**0,07****
2011 - 2010	0,00	-0,01	0,00	0,00
2012 - 2011	**0,07****	**0,07****	**0,05****	**0,06****
2012 - 2008	**0,11****	**0,14****	**0,12****	**0,12****

* a < ,01 **: a < ,001

Tabel 2.5 Vergelijking over de jaren heen in percentage per leerlingengroep voor de rekenonderwerpen in jaargroep 8

	Percentiel-waarde	Getallen en bewerkingen				Breuken, procenten en verhoudingen				Meten, meetkunde, tijd en geld				
Leerlingengroep*	2008	09	10	11	12	09	10	11	12	09	10	11	12a	12b
> dan ZLV	**10**	90	91	91	92	90	91	91	92	90	91	91	92	90
> LV	**25**	75	76	76	78	75	77	77	79	74	77	77	79	75
> Standaard 2008	**50**	50	52	51	54	50	53	53	56	49	53	53	55	50
> HV	**75**	25	26	26	28	25	27	27	30	24	28	27	29	25
>ZHV	**90**	10	11	11	12	10	11	11	13	10	11	11	12	10

* zie tabel 1.2

2.2 Verschillen voor achtergrondvariabelen

In deze paragraaf rapporteren we de verschillen tussen groepen van leerlingen gebaseerd op een categorisering naar achtergrondvariabele. We presenteren hier in de eerste plaats de gevonden effecten in 2012 en gaan bij enkele variabelen in op de trend over vijf jaar.

Het gebruikte basismodel voor de analyses van 2012 bevatte de verklarende variabelen geslacht, leertijd, formatiegewicht en stratum. Ieder van de geschatte effecten per variabelen wordt gecorrigeerd voor effecten van de overige variabelen. Hiermee kan het unieke effect van de variabele bepaald worden.
Aan dit basismodel zijn in de verdere analyses variabelen toegevoegd. Dit betreft de variabelen toetsvariant (Eindtoets of Niveautoets), IJK (“speciale categorie leerling”), advies VO, schoolgrootte, regio en verstedelijking. Dit levert een vijftal extra modellen op de basisvariabelen telkens aangevuld met één extra variabele. De gerapporteerde effecten zijn telkens gecorrigeerde effecten. Voor een weergave van de gemiddelden, standaarddeviaties, overschrijdingskansen en effectgroottes per achtergrondvariabele verwijzen we naar de tabellen in de bijlagen. In tabel 2.6 zijn de effectgroottes voor de variabelen op leerlingniveau weergegeven.

*Tabel 2.6 Effecten op leerlingniveau in jaargroep 8 voor het contrast 2012-2011**

Variabele	Contrast	Taal			Rekenen		
		WS	Sp	BL	G/B	BPV	MMTG
Geslacht	Meisjes - Jongens	0,00	**0,25**	0,12	**-0,43**	**-0,43**	**-0,39**
Leertijd	Vertraagd - Regulier	**0,50**	**0,69**	**0,60**	**0,67**	**0,71**	**0,71**
Formatiegewicht	0.0 - 0.3	**0,48**	**0,33**	**0,59**	**0,49**	**0,56**	**0,56**
	0.0 - 1.2	**0,76**	0,06	**0,57**	**0,27**	**0,36**	**0,43**
	0.3 - 1.2	**0,28**	**-0,27**	-0,03	-0,23	-0,20	-0,12
Advies VO	vmbo-KB - vmbo-BB	**0,80**	**0,54**	**1,09**	**1,17**	**1,14**	**1,13**
	vmbo-GT - vmbo-KB	**0,47**	**0,42**	**0,71**	**0,65**	**0,72**	**0,67**
	havo - vmbo-GT	**0,93**	**0,90**	**1,34**	**1,20**	**1,32**	**1,36**
	vwo - havo	**1,34**	**1,51**	**1,52**	**1,44**	**1,61**	**1,83**
	vwo - vmbo BB	**3,53**	**3,38**	**4,66**	**4,46**	**4,79**	**4,99**
Toets	Eindtoets-Niveautoets	**1,10**	**0,79**	**1,45**	**1,86**	**1,62**	**1,71**

* vet is significant (p=0.01)

2.2.1 Effecten op leerlingniveau

Geslacht

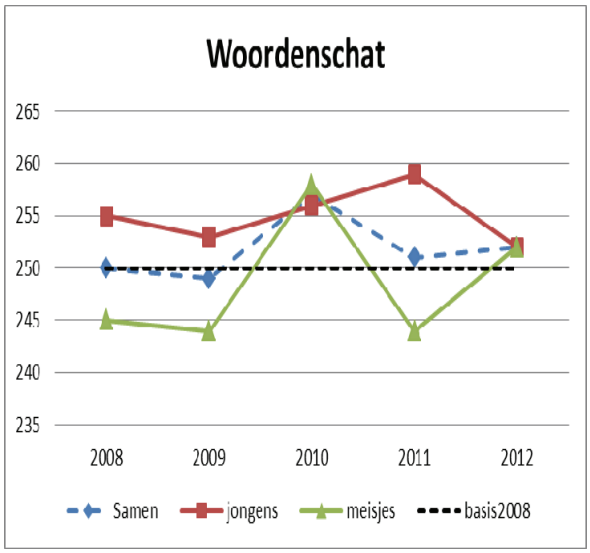
Voor het verschil in resultaten tussen jongens en meisjes zien we in veel gevallen significante en betekenisvolle effecten. Deze zijn bij de rekenonderwerpen het grootst en in het voordeel van de jongens. Bij Woordenschat is er geen verschil en het gevonden verschil bij Begrijpend lezen is niet significant. Spelling laat wel een significant verschil zien in het voordeel van de meisjes. De effectgrootte varieert van klein tot matig.

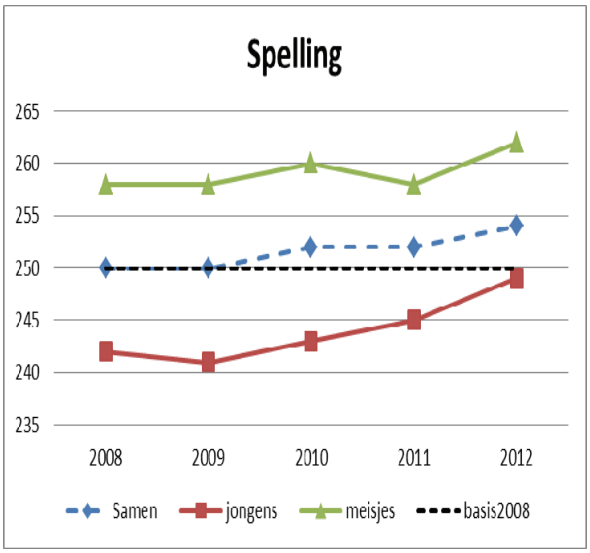
In tabel 2.7 hebben we deze resultaten onderscheiden naar geslacht bij elkaar gezet, verder in beeld gebracht met een grafische weergave per vaardigheid in figuur 2.4. Doorgaans zien we parallelle ontwikkelingen, maar Woordenschat is daarop een uitzondering. Met name de resultaten van de meisjes zijn grillig te noemen. De lagere resultaten van de jongens in 2012 worden meer dan gecompenseerd door de positievere resultaten bij de meisjes, zodat de algemene richting positief is. Bij de overige vaardigheden zien we wel steeds een min of meer parallelle trendlijn. Ook valt op dat de winst bij de meisjes sinds 2008 bij twee van de drie rekenonderwerpen groter is dan bij de jongens.

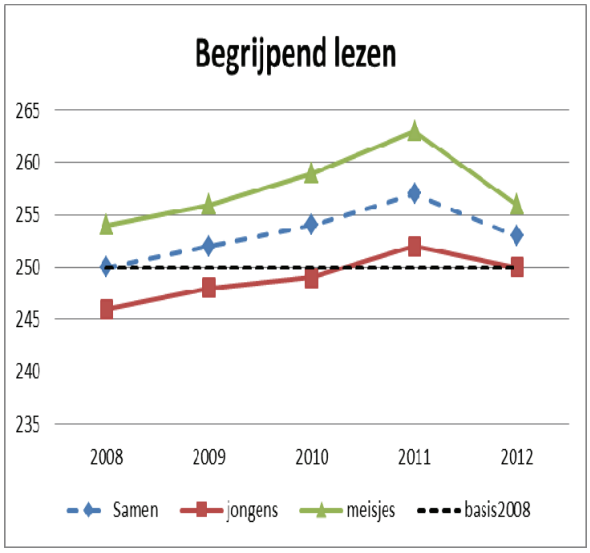
Tabel 2.7 Vaardigheidsscores per onderwerp naar geslacht van 2008 tot en met 2012

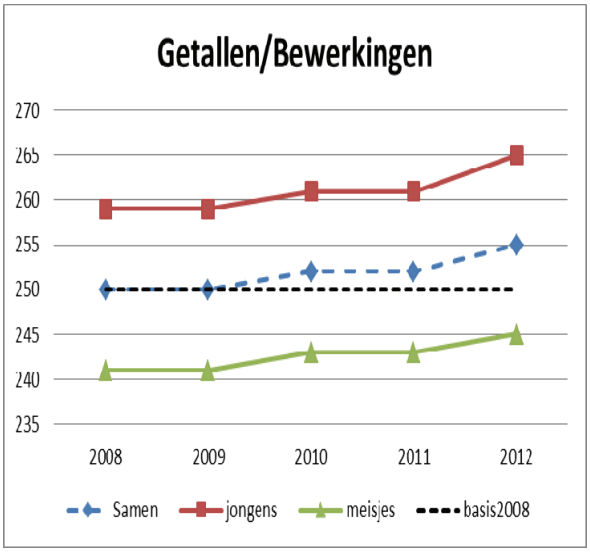
	Geslacht	Gemiddelde vaardigheidsscore					Verschil
		2008	2009	2010	2011	2012	2012-2008
Taal							
Woordenschat	samen	250	249	257	251	252	2
	jongens	255	253	256	259	252	-3
	meisjes	245	244	258	244	252	7
Spelling	samen	250	250	252	252	254	4
	jongens	242	241	243	245	249	7
	meisjes	258	258	260	258	262	4
Begrijpend lezen	samen	250	252	254	257	253	3
	jongens	246	248	249	252	250	4
	meisjes	254	256	259	263	256	2
Rekenen-Wiskunde							
Getallen/Bewerkingen	samen	250	250	252	252	255	5
	jongens	259	259	261	261	265	6
	meisjes	241	241	243	243	245	4
Breuken, procenten, verhoudingen	samen	250	250	254	253	257	7
	jongens	261	261	266	264	266	5
	meisjes	239	239	241	243	248	9
Meetkunde, meten, tijd en geld	samen	250	249	254	254	256	6
	jongens	261	260	265	266	265	4
	meisjes	239	239	243	242	248	9

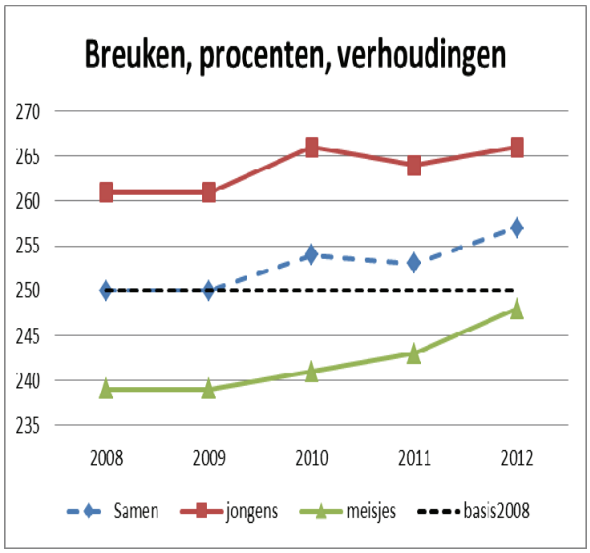
Figuur 2.4 Trends in vaardigheden naar geslacht

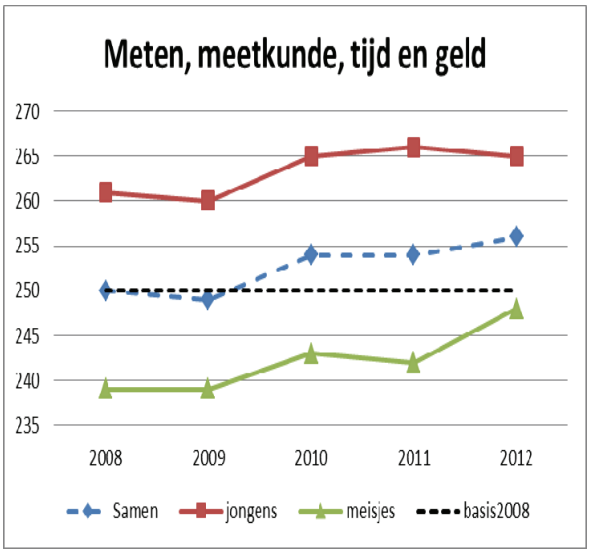












Leertijd
Het verschil in resultaat ten gevolge van de categorisering naar leertijd is overal significant en betekenisvol. De effectgrootte is overal matig.

Formatiegewicht
Voor de variabele formatiegewicht vinden we niet overal significante effecten. Met name het contrast tussen leerlingen met formatiegewicht 0.3 en leerlingen met formatiegewicht 1.2 is vaak niet significant. Alleen voor de taalonderwerpen Woordenschat en Spelling vinden we daar significante verschillen. Opvallend is hier dat deze effecten verschillende kanten op wijzen: leerlingen met een hoog gewicht scoren beter in Spelling, maar slechter in Woordenschat, dan leerlingen met een laag gewicht. Bij het contrast tussen leerlingen met een hoog gewicht en leerlingen zonder gewicht zien we matige effectgroottes ten gunste van de leerlingen zonder gewicht, uitgezonderd bij Spelling. Ook het contrast tussen leerlingen met een laag gewicht en leerlingen zonder gewicht (0.00 versus 0.30) is steeds significant, met een effectgrootte die varieert van klein tot matig.

VO-advies
We zien voor de variabele VO-advies dat er sprake is van contrasten die significant zijn en betekenisvolle effectgroottes opleveren. Die effectgroottes variëren van klein tot (zeer) groot. Bij het grootste contrast, dat tussen vwo en vmbo basis, zien we effectgroottes tot 4,99, hetgeen betekent dat het verschil tussen deze groepen van leerlingen, gecorrigeerd voor geslacht, leertijd, en formatiegewicht, bijna vijf standaard-deviaties bedraagt. Het illustreert de grote spreiding in vaardigheid die we in jaargroep 8 al aantreffen. Opvallend in dit verband is ook de aanzienlijke effectgrootte bij het contrast vwo-havo. Deze leerlingen zitten in het VO soms nog bij elkaar in een klas, maar ook hier is met name bij het rekenonderdeel Meetkunde, meten, tijd en geld het verschil aanzienlijk.

Toets
Ook de toetsvorm die is gemaakt kan als variabele worden gezien. Leerlingen die een achterstand hebben van anderhalf jaar of meer krijgen van de leerkracht de niveautoets toegewezen. Dit is uiteraard een naar verhouding kleine groep leerlingen. Het te verwachten verschil in prestatie zien we terug in de effect-groottes. Die variëren van 0,79 bij Spelling tot 1,96 bij de schaal Getallen/Bewerkingen bij Rekenen-Wiskunde en zijn dus allen als groot te kwalificeren, met als enige uitzondering Spelling. Daar zien we een lagere effectgrootte, die als matig gekwalificeerd moet worden.

De variabele IJK
In de Eindtoets van 2012 is bij het registreren van achtergrondvariabelen van de leerling helaas niet zoals in vorige jaren een code opgenomen voor de thuistaal van de leerlingen. Wel is net als in eerdere jaren de mogelijkheid gebleven voor de leerkracht om aan te kruisen of de leerling in een speciale categorie thuishoort[1]:

Code I: (Allochtone) leerlingen die aan het begin van jaargroep 8 vier jaar of korter in Nederland zijn en die het Nederlands onvoldoende beheersen om de opgaven in de Eindtoets goed te kunnen lezen.
Code J: Leerlingen die naar verwachting naar het (voortgezet) speciaal onderwijs of naar het praktijkonderwijs (pro) gaan
Code K: Leerlingen die naar verwachting in aanmerking komen voor het leerwegondersteunend onderwijs (lwoo).

[1] Of de leerkracht deze codes op de juiste wijze heeft gebruikt is niet of maar beperkt te controleren.

Tabel 2.8 Effecten van de variabele IJK en gemiddelden en standaarddeviaties voor 2012 en 2011

		2012			2011		
		Effect	**M**	**SD**	**Effect**	**M**	**SD**
T: Spelling	Geen IJK	0	260	48	0	255	48
	Wel IJK	**-0,807**	213	49	**-0,977**	208	48
T: Begrijpend lezen (anker)	Geen IJK	0	258	47	0	262	47
	Wel IJK	**-1,148**	193	49	**-1,340**	198	47
T: Begrijpend lezen (basisvaardigheden)	Geen IJK	0	267	47	0	262	47
	Wel IJK	**-1,207**	200	48	**-1,148**	199	47
T: Woordenschat	Geen IJK	0	256	48	0	255	48
	Wel IJK	**-0,915**	201	50	**-1,100**	202	48
RW: Getallen en bewerkingen	Geen IJK	0	260	47	0	257	47
	Wel IJK	**-1,250**	194	48	**-1,371**	192	47
RW: Breuken, procenten en verhoudingen	Geen IJK	0	262	46	0	257	47
	Wel IJK	**-1,219**	197	47	**-1,266**	197	47
RW: Meten, tijd en geld	Geen IJK	0	261	47	0	258	48
	Wel IJK	**-1,161**	197	48	**-1,256**	199	48

***** Alle verschillen zijn significant (p=0.01)

Leerlingen in de categorieën I en J doen doorgaans niet mee aan de Eindtoets, maar kunnen wel de Niveautoets voorgelegd krijgen. Voor leerlingen in de categorie K is dat zeker het geval. Daarvan vinden we er ook het meest terug in de data. Er is een grote samenhang tussen formatiegewicht en deze speciale codes. Leerlingen met de code K komen het procentueel het meest voor bij successievelijk 1.20- en 0.30-leerlingen (18% en 15% bij gemiddeld 7,9%). Dat geldt ook voor de codes I en J, maar dan met veel lagere percentages (ca. 2-1 %).

In tabel 2.8 zien we dat bij alle onderwerpen het contrast tussen leerlingen die wel een IJK-code hebben en leerlingen die dat niet hebben significant en betekenisvol is. De effectgroottes zijn zonder uitzondering groot en hebben in de meeste gevallen de omvang van meer dan een hele standaarddeviatie. Alleen bij Spelling is het effect iets kleiner, maar nog steeds groot te noemen (>0.80). Voor de volledigheid zijn ook de effecten van 2011 in kaart gebracht. Daar bleken de effectgroottes nog iets forser te zijn, uitgezonderd Begrijpend lezen (toets Basisvaardigheden).

2.2.2 Effecten op schoolniveau

Bij de analyse van de invloed van de achtergrondvariabelen stratum, schoolgrootte, regio en verstedelijking is ook een correctie toegepast. Bij de variabelen is namelijk eveneens een duidelijk een relatie met andere variabelen die maken dat voor een goede vergelijking een gezuiverd effect moet worden berekend. Daarin zijn de variabelen geslacht, leertijd en formatiegewicht als correctiefactoren meegenomen (zie tabel 2.9). De indeling van scholen naar stratum is gebaseerd op dezelfde variabele als formatiegewicht en we zien dat in de effecten bij de contrasten terug. Het beeld is vergelijkbaar, alleen is het aantal significante verschillen en betekenisvolle effecten aanzienlijk kleiner. Bij Spelling en bij Rekenen: Getallen en bewerkingen doen ze zich niet voor. Bij de overige twee rekenonderwerpen is alleen het contrast ‘weinig vs. veel’-gewichtenleerlingen betekenisvol. Bij de taalonderwerpen Woordenschat en Begrijpend lezen levert ook het contrast ‘matig versus veel gewichtenleerlingen nog een betekenisvol effect op.
Voor de overige variabelen blijkt na correctie weinig verschil over te blijven. De schoolvariabelen grootte, regio en verstedelijking doen er zo bezien weinig toe. Opgemerkt kan worden dat als naar de gemiddelde vaardigheden gekeken wordt, waarbij er niet gecorrigeerd wordt voor geslacht, leertijd en formatiegewicht, de leerlingen in de zeer sterk verstedelijkte gebieden wel wat lager scoren. Dat zal vooral het gevolg zijn van het grotere percentage achterstandsleerlingen in de grote steden.

Tabel 2.9 Effecten van de variabelen stratum, schoolgrootte, regio en verstedelijking (na correctie)

Variabele	Contrast	Taal			Rekenen		
		WS	Sp	BL	G/B	BPV	MMTG
Stratum (% gew. lln.)*	Weinig - Matig	0,12	0,07	0,17	0,09	0,12	0,12
	Matig - Veel	**0,28**	0,05	**0,29**	0,04	0,11	0,15
	Weinig - Veel	**0,40**	0,12	**0,45**	0,14	**0,24**	**0,27**
Schoolgrootte	Groot - Klein	-0,03	0,13	0,07	0,08	0,07	0,05
Regio	Oost - Noord	-0,05	-0,01	0,10	0,04	0,06	0,06
	West - Noord	0,00	0,06	0,13	0,09	0,09	0,09
	Zuid - Noord	-0,07	0,11	0,14	0,11	0,09	0,10
	West - Oost	0,05	0,07	0,03	0,05	0,02	0,03
	Zuid - Oost	-0,02	0,12	0,04	0,06	0,03	0,04
	Zuid - West	-0,08	0,05	0,01	0,01	0,01	0,01
Verstedelijking	Sterk - Zeer sterk	-0,01	-0,11	-0,01	-0,03	-0,04	-0,04
	Matig - Sterk	0,01	0,01	0,02	0,00	0,01	0,03
	Weinig - Matig	-0,01	-0,05	-0,02	-0,03	-0,02	-0,05
	Niet - Weinig	0,00	-0,06	-0,04	-0,05	-0,05	0,01

* vet is significant (p=0.01); Stratumindeling naar % gewichtsleerlingen: <10% = weinig, 10 – 25%= matig, > 25% = veel.

3 De resultaten voor jaargroep 4

In de peiling van 2011 is voor het eerste gerapporteerd op een algemene rekenschaal. Dat was mogelijk omdat de subschalen voor Rekenen zeer hoog met elkaar correleerden. Hierdoor leken ook de resultaten voor de verschillende subschalen sterk op elkaar: de belangrijkste informatie werd dan ook goed weergegeven door de algemene rekenschaal zonder veel verlies aan informatie. In de dataverzameling is er dan ook voor gekozen om de totaalscores voor Rekenen, zonder de scores op de subschalen, te verzamelen.

3.1 De vergelijking over de jaren

Bij jaargroep 4 zien we in de resultaten een beeld dat sterk overeenkomt met dat van jaargroep 8. Bij alle variabelen is de score in schaalwaarde hoger dan in 2008 (zie tabel 3.1). Bij Woordenschat zijn de metingen met het betreffende instrument pas in 2009 gestart en ook daar is de score duidelijk omhoog gegaan.

Tabel 3.1 Effectgroottes en gemiddelde schaalwaarden voor taal en rekenen jaargroep 4

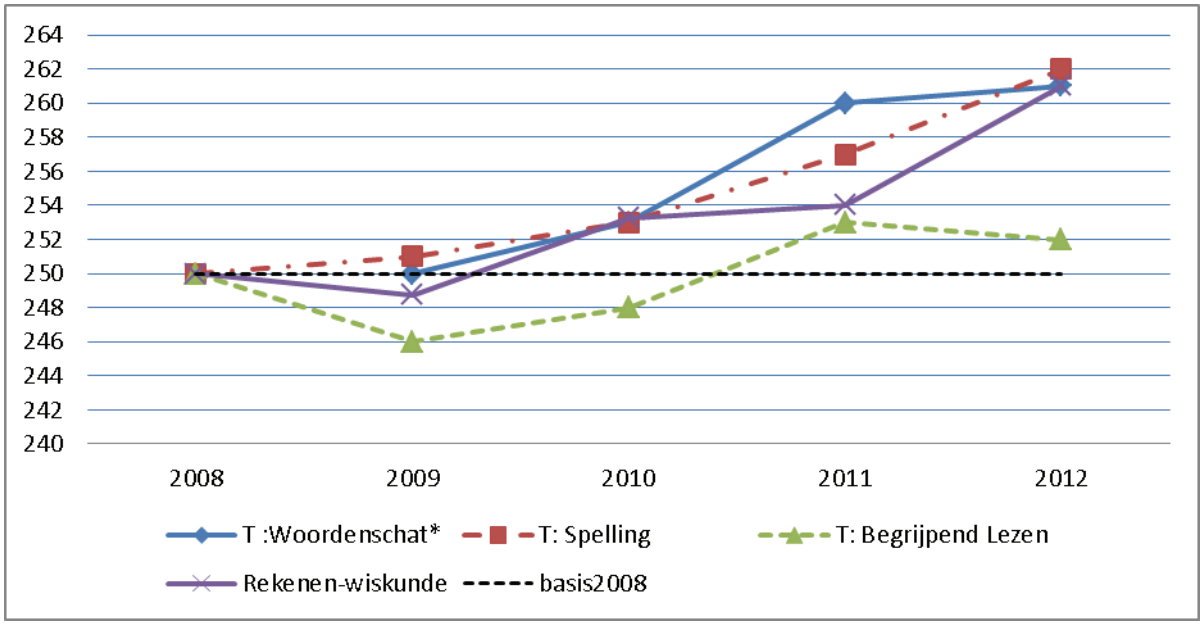
Onderwerpen:	**Effecten***		**Jaren**				
	2012-2011	**2012-2008**	**2008**	**2009**	**2010**	**2011**	**2012**
T: Woordenschat*	0,03	**0,21**	-	250	253	260	261
T: Spelling	**0,15**	**0,24**	250	251	253	257	264
T: Begrijpend Lezen	-0,02	0,04	250	246	248	253	252
RW: Rekenen-Wiskunde	**0,12**	**0,22**	250	249	255	254	261

* Vet =significant, p=0,01; Woordenschat is vanaf 2009 gemeten, contrast is hier 2012-2009

Evenals in het voorgaande jaar valt op dat niet overal een plus is te melden als we een vergelijking maken met het voorafgaande jaar. Bij Begrijpend lezen is de score een punt gezakt, maar deze daling is niet significant. Bij Rekenen-wiskunde, waar in 2011 ook een teruggang van een punt werd waargenomen, is nu echter een vooruitgang te melden van 7 punten. Omdat nu geen onderscheid meer is gemaakt tussen de vier onderwerpen bij Rekenen is niet aan te geven of de vooruitgang algemeen is of per onderwerp verschilt. De andere twee taalonderwerpen, Woordenschat en Spelling, laten beiden een positief verschil zien vergeleken met 2011, maar het verschil bij Woordenschat is gering en niet significant. De stijging van 7 punten bij Spelling is wel significant.

De trends voor de vier vaardigheden zijn in figuur 3.1 grafisch weergegeven.

Figuur 3.1 Trends voor jaargroep 4



Met behulp van de percentages in tabel 3.2 kunnen we een inschatting maken van de omvang van de verschillen. Zo zien we dat bij Woordenschat in 2012 59% van de leerlingen de gemiddelde score van 2008 behaalt. Er is dus sprake van een toename van 9%. Voor Spelling is dat ook 9% en bij Begrijpend lezen is dat 2%. Voor de rekenvaardigheid, is nu een gemeenschappelijke schaal gemaakt, waarin de vier onderdelen zijn gecombineerd. De resultaten van de voorliggende jaren, 2009 tot en met 2011, zijn verkregen door de gemiddelden over de drie eerder gerapporteerde rekenschalen te nemen. Hier zien we ook een toename van het percentage leerlingen dat de P50 van 2008 behaald. Dat is nu ook 59%, een toename van 9%.
We zien dus bij alle onderwerpen dat meer dan de helft van de leerlingen in 2012 boven het gemiddelde van 2008 (de standaard) presteert.

Tabel 3.2 Vergelijking over de jaren heen in percentage per leerlingengroep voor jaargroep 4

		Woordenschat				**Spelling**				**Begrijpend lezen**				**Rekenen-wiskunde**			
Leerlingengroep*	**2008**	**'09**	**'10**	**'11**	**'12**	**'09**	**'10**	**'11**	**'12**	**'09**	**'10**	**'11**	**'12**	**'09**	**'10**	**'11**	**'12**
> dan ZLV	**10**	90	91	93	93	90	92	92	94	89	89	91	91	90	91	91	94
> LV	**25**	75	77	81	81	76	77	79	82	72	74	77	76	74	77	78	83
> Standaard 2008	**50**	50	53	58	59	51	53	56	59	47	48	52	52	49	52	54	61
> HV	**75**	25	27	32	32	26	27	30	33	23	23	27	26	24	27	28	34
>ZHV	**90**	10	11	14	14	10	11	13	15	9	9	11	11	10	11	12	15

* zie tabel 1.2

Behalve als verschillen op de schaal kunnen de jaarverschillen ook als effecten weergegeven worden. Bij elkaar zien we in tabel 3.3 dat bij de vergelijking van 2012 met 2011 drie van de vier schalen een verschil opleveren in positieve richting en één in negatieve richting ten opzichte van 2011. Van deze verschillen was de stijging bij Spelling en bij Rekenen significant, maar te klein om te voldoen aan de kwalificatie betekenisvol (ze zijn kleiner dan .20). Vergelijken we de uitkomsten over de hele periode van 2008 tot en met die van 2012, dan is er bij drie onderwerpen sprake van een significant verschil in positieve richting en dat is in alle drie gevallen een betekenisvol effect. Begrijpend lezen laat over de jaren heen de minste vooruitgang zien. De effectgroottes geven aan dat het verschil in schaalwaarde tussen 2008 en 2012 bij Rekenen, Spelling en Woordenschat significant is, maar bij Begrijpend lezen niet.

Tabel 3.3 Effectgroottes voor de jaarvergelijkingen per onderwerp (gecorrigeerd)

Contrasten	Taal			Rekenen-Wiskunde
	Woordenschat	Spelling	Begrijpend lezen	totaal
2009 – 2008	n.v.t.	0,07	-0,05	-0,01
2010 – 2009	0,04	0,00	0,00	0,06
2011 – 2010	0,14**	0,08*	0,11**	0,04
2012 – 2011	0,03	0,15**	-0,03	0,12**
2012 – 2008	0,21**	0,30**	0,04	0,21**

Significantie:* a < ,01 ; **: a < ,001; Rekenen-Wiskunde betreft effectgrootte van 4 onderdelen gezamenlijk

3.2 Verschillen per achtergrondvariabele

In deze paragraaf gaan we meer in detail in op de prestatieverschillen tussen vierdegroepers met een verschillende achtergrond. We presenteren voor 2012 de verschillen voor de achtergrondvariabelen van de leerlingen, nl. geslacht, leertijd, formatiegewicht en thuistaal. Op schoolniveau rapporteren we over de achtergrondvariabelen stratum, regio en verstedelijking (urbanisatiegraad). Ook hier moet bedacht worden dat niet alle in statistisch opzicht significante contrasten ook betekenisvol zijn (zie paragraaf 3.1).
De volgende jaarlijkse peilingen moeten uitwijzen in hoeverre de hier geconstateerde verschillen stabiel zijn over de jaren heen. De hier gerapporteerde effecten betreffen, net als in eerdere rapportages, de gecorrigeerde variant, waarbij de (basis)variabelen geslacht, leertijd, formatiegewicht en stratum steeds zijn meegenomen in het model. De effecten voor deze variabelen betreffen de effecten zoals gevonden zijn in dit basismodel. Ook weer vergelijkbaar met eerdere rapportages geldt voor de overige extra variabelen dat de effecten gerapporteerd worden zoals gevonden in het model met de basisvariabelen waarbij alleen die extra variabele is toegevoegd. We rapporteren deze effecten gescheiden naar variabelen op leerlingniveau en variabelen op schoolniveau, hoewel dat onderscheid in de analyse verder geen rol speelt.

Voordat we de verschillen in effecten bespreken kunnen we al constateren dat er meer significante verschillen en betekenisvolle effecten zijn bij de leerlingvariabelen dan bij de schoolvariabelen. Variabelen als regio en verstedelijking blijken weinig verschil in resultaten op te leveren als gecorrigeerd wordt voor de basisvariabelen. Bij de achtergrondvariabelen voor de leerlingen ligt dat anders. Daar vinden we veel kleine en enkele matige effecten en één keer een groot effect. Bij de schoolvariabelen zien we alleen kleine effecten.

3.2.1 Effecten van leerlingvariabelen

De contrasten bij de leerlingvariabelen leveren in veel gevallen significantie op en de effectgrootte is vaak betekenisvol (zie tabel 3.4). We bespreken de uitkomst van deze analyses per variabele.

Tabel 3.4 Effectgrootten van achtergrondvariabelen op leerlingenniveau voor 2012 in jaargroep 4 (gecorrigeerd)*

Variabele	Contrast	Rekenen-Wiskunde	Woordenschat	Begrijpend lezen	Spelling
Geslacht	Jongens-meisjes	**0,36**	*-0,10*	**-0,21**	**-0,19**
Leertijd	Regulier-Vertraagd	**0,34**	**0,24**	**0,30**	**0,38**
Formatiegewicht	0.0 - 0.3	**0,41**	**0,34**	**0,32**	**0,30**
	0.0 - 1.2	**0,40**	**0,51**	**0,41**	0,11
	0.3 - 1.2	0,00	0,18	0,09	-0,19
Herkomst	Nederland-elders	*0,24*	**0,30**	*0,23*	**-0,42**
Thuistaal	Alleen Nls - Nls & Bui	**0,54**	**0,69**	**0,44**	0,10
	Alleen Nls - Alleen Bui	**0,45**	**0,90**	**0,51**	0,11
	NLs & Bui - Alleen Bui	-0,09	0,21	0,07	0,01

* vet= significant (p<0.01); cursief =significant (p<0.05)

Geslacht

Jongens behalen over het algemeen lagere scores voor de taalvaardigheden dan meisjes en dat verschil is in 2012 voor alle drie onderwerpen significant. Het verschil bij Woordenschat is het kleinst en is net als in jaargroep 8 verwaarloosbaar. De andere contrasten leveren kleine effectgroottes op. Bij Rekenen is de situatie anders: daar is het contrast significant en betekenisvol in het voordeel van de jongens. Ook hier is de effectgrootte als klein te kwalificeren en ook iets kleiner dan in 2011 (zie tabel 3.5).

Als we de resultaten op de vaardigheidsschaal[2] voor jongens en meisjes bekijken in de trendlijnen over de afgelopen vijf jaren, dan valt op dat die lijnen meestal parallel lopen. Alleen bij Woordenschat zien we een afwijking van dat beeld: de meisjes gaan vooruit, maar de jongens niet. Verder is duidelijk dat bij de taalonderdelen de meisjes steeds hoger scoren dan de jongens, maar bij Rekenen-Wiskunde is de situatie andersom. Over de afgelopen vijf jaar bezien maken de meisjes de meeste progressie bij Woordenschat en de jongens bij Spelling. De vooruitgang bij Rekenen-Wiskunde is niet verschillend voor jongens en meisjes en bij Begrijpend lezen is er geen vooruitgang en geen onderscheid daarin. Deze trends zijn weergegeven in afzonderlijke figuren (figuur 3.2).

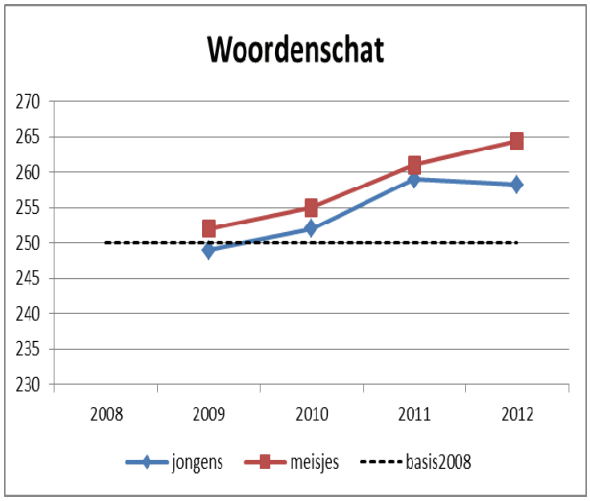
[2] Resultaten op de vaardigheidsschaal zijn de resultaten waarbij niet gecorrigeerd wordt voor de effecten van andere variabelen.

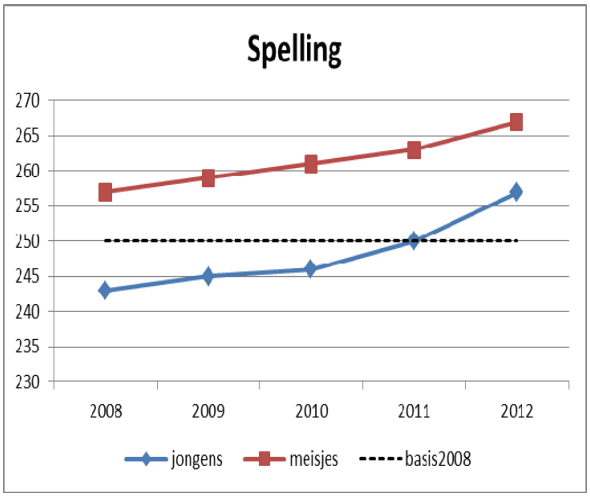
Tabel 3.5 Trend in schaalwaarden van 2008 tot en met 2012 naar geslacht

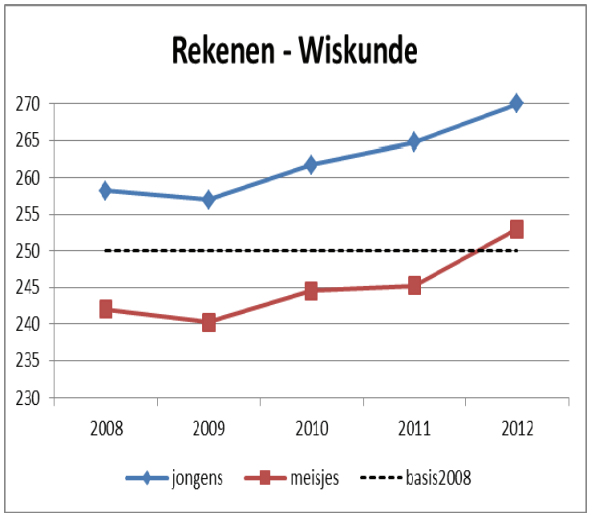
Onderwerpen:	Gemiddelde vaardigheidsscores					toename
	2008	2009	2010	2011	2012	2012-2008*
Woordenschat*	-	250	253	260	261	11
Jongens	-	249	252	259	258	9
Meisjes	-	252	255	261	264	12
Spelling	250	251	253	257	264	14
Jongens	243	245	246	250	259	16
Meisjes	257	259	261	263	269	12
Begrijpend Lezen	250	246	248	253	252	2
Jongens	245	242	243	248	247	2
meisjes	255	251	253	258	257	2
Rekenen-Wiskunde	250	249	253	254	261	11
jongens	258	257	262	265	270	12
meisjes	242	240	245	244	253	11

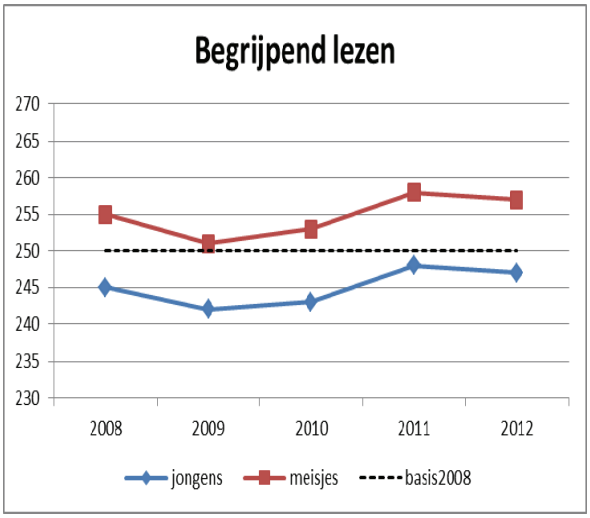
* Start van de meting Woordenschat in 2009

Figuur 3.2 Trends in vaardigheid naar geslacht









Leertijd
Vertraagde leerlingen, d.w.z. degenen die ten minste een keer doubleerden, behalen aanzienlijk lagere scores dan niet-vertraagde leerlingen. Dit contrast levert over de hele linie een klein effect op. Net als in 2011 was het effect het kleinst bij Woordenschat en het grootst bij Spelling.

Formatiegewicht
Bij de start van het project (2008) rapporteerden scholen (ook) nog oude formatiegewichten, maar de laatste jaren hebben we alleen te maken met de nieuwe formatiegewichten. Bij dit achtergrondkenmerk zien we een gevarieerd beeld. Voor Rekenen-Wiskunde zijn de contrasten tussen leerlingen met een gewicht, laag of hoog, in vergelijking met leerlingen zonder formatiegewicht significant. De bijbehorende effectgrootte moet als matig gekwalificeerd worden, terwijl het contrast laag versus hoog formatiegewicht geen significant verschil oplevert. Dit resultaat is vergelijkbaar met het resultaat in 2011.
Bij de taalvaardigheden zijn er kleine verschillen waar te nemen met 2011, maar in grote lijnen is het resultaat vergelijkbaar. Bij alle taalvaardigheden zien we weer een significant klein effect tussen de leerlingen zonder formatiegewicht (0.00) en de leerlingen met een klein formatiegewicht (0.30).
De ongecorrigeerde verschillen zijn nog iets groter. De verschillen zijn in vergelijking met 2011 iets kleiner geworden voor Begrijpend lezen en Woordenschat, en iets groter voor Spelling. De verschillen tussen de leerlingen met een hoog formatiegewicht (1.20) en de leerlingen met een laag formatiegewicht (0.30) wisselen per vaardigheid. Bij Begrijpend lezen is er net als bij Rekenen ook amper een verschil tussen leerlingen met een hoog en een laag formatiegewicht. Bij Woordenschat presteren de leerlingen met een laag formatiegewicht iets beter en bij Spelling de leerlingen met een hoog formatie. Deze effecten zijn echter niet significant en ook niet als betekenisvol te classificeren.
Voor de verschillen tussen leerlingen zonder gewicht en met een hoog formatiegewicht betekent het dat voor alle vaardigheden behalve Spelling er een significant effect is. In vergelijking met 2011 was in 2012 bij alle taalvaardigheden het verschil tussen 0.00- en 1.20-leerlingen kleiner. Bij Rekenen was het verschil in beide jaren gelijk.

Als de groepen op de vaardigheidsschaal vergeleken worden zijn de verschillen nog wat groter dan wanneer er niet gecorrigeerd wordt voor het effect van formatiegewicht en leertijd.

Herkomst
Deze variabele laat significante verschillen zien voor alle onderwerpen. Spelling valt daarbij op omdat het effect omgekeerd is aan dat bij de andere drie onderwerpen: De leerlingen met een niet-Nederlandse herkomst behalen in dit geval hogere scores. De effectgrootte moet overal als matig worden gekwalificeerd.

Thuistaal[3]
Voor deze variabele hebben we de resultaten van de meting ook over de jaren heen per categorie in kaart gebracht. We zien in tabel 3.6 dat de ontwikkeling per groep duidelijk verschilt. De algemene trend die we in figuur 3.1 al zagen, zien we hier uiteraard terug bij de leerlingen die thuis alleen Nederlands spreken. Dat is de grootste groep leerlingen. Bij de andere twee categorieën zijn de trends echter meer gevarieerd.
De trends per categorie zijn weergegeven in afzonderlijke figuren (figuur 3.3).

Het is opvallend dat de leerlingen die naast Nederlands een andere taal spreken op alle vaardigheden minder goed presteren dan in 2011. Deze groep leerlingen was ook in vergelijking met andere jaren een minder goede groep, terwijl de groep in 2011 juist zeer sterk presteerde. De gevonden gecorrigeerde effecten bij de variabele thuistaal in 2012 zaten ook relatief dichter bij gecorrigeerde effecten van de peilingen voor 2011. Bij de groep leerlingen die thuis alleen een andere taal dan Nederlands spreekt was het resultaat alleen bij Woordenschat iets slechter dan in 2011. Bij de overige vaardigheden bleef het niveau constant.

---

[3] In deze variabele is geen onderscheid gemaakt naar buitenlandse taal. In de meeste gevallen gaat het om Marokkaans, Turks en Surinaams/Antilliaans.

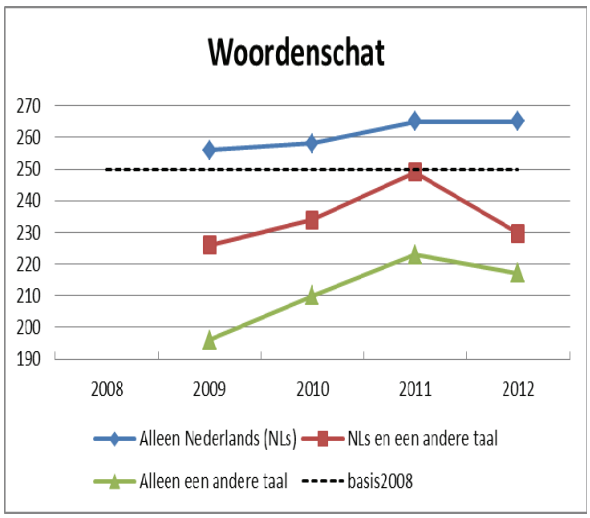
Meest opvallend is de consistente daling voor de groep leerlingen met een gemengde thuistaal, terwijl de score bij de groep leerlingen met alleen een buitenlandse thuistaal constant blijft. Daardoor liggen de niveaus van deze twee groepen weer dichter bij elkaar en soms zo dicht bij elkaar dat ze gelijk zijn.

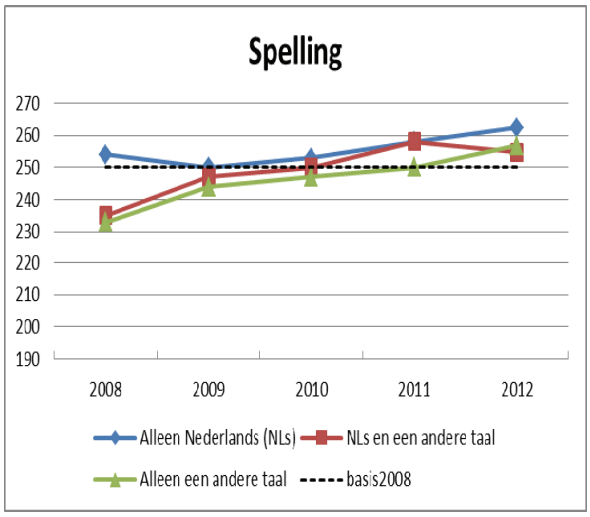
Tabel 3.6 Trend in schaalwaarden van 2008 tot en met 2012 voor de categorieën van thuistaal

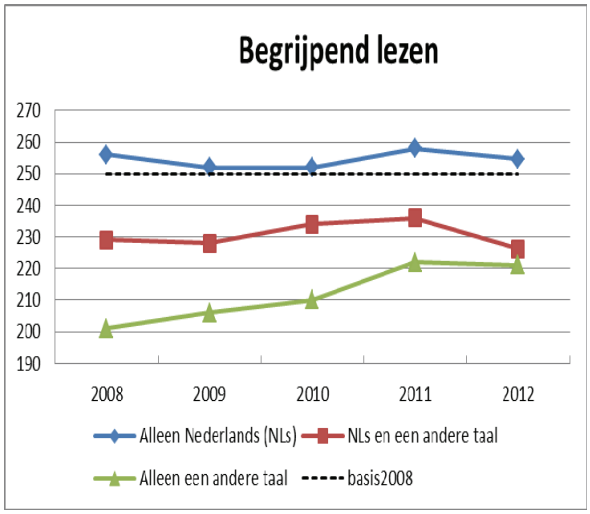
Onderwerp	Categorie thuistaal	2008	2009	2010	2011	2012	2012-2008
Woordenschat*	Alleen Nederlands (NLs)	-	256	258	265	265	9
	NLs en een andere taal	-	226	234	249	230	4
	Alleen een andere taal	-	196	210	223	217	21
Spelling	Alleen Nederlands (NLs)	254	250	253	258	265	11
	NLs en een andere taal	235	247	250	258	260	25
	Alleen een andere taal	233	244	247	250	257	24
Begrijpend lezen	Alleen Nederlands (NLs)	256	252	252	258	255	-1
	NLs en een andere taal	229	228	234	236	226	-3
	Alleen een andere taal	201	206	210	222	221	20
Rekenen-Wiskunde	Alleen Nederlands (NLs)	256	253	257	257	264	8
	NLs en een andere taal	224	236	238	252	230	6
	Alleen een andere taal	216	220	223	234	232	16
Gemiddelde trend	Alleen Nederlands (NLs)	256	253	256	258	262	0
	NLs en een andere taal	227	235	238	250	235	10
	Alleen een andere taal	216	218	223	233	232	21

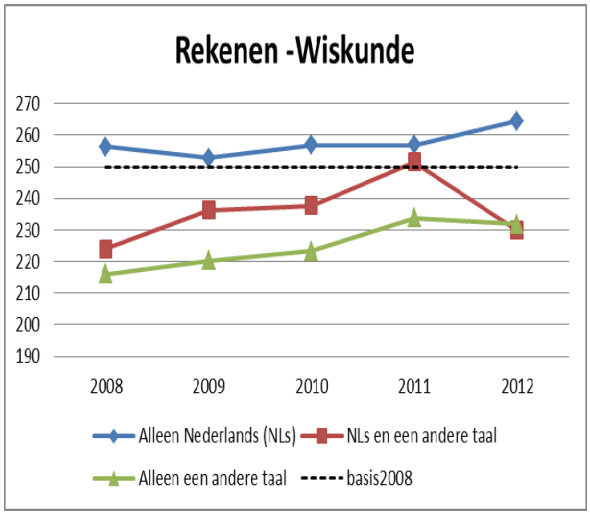
* Start van de meting in 2009

Figuur 3.3 Trends in vaardigheid naar thuistaal









3.2.2 Effecten van schoolvariabelen

Stratum

Het contrast op basis van de indeling in strata levert bij alle vaardigheden significante verschillen op. Bij Rekenen, Woordenschat en Begrijpend lezen is te zien dat de scholen met relatief veel achterstands-leerlingen (>25%; stratum 3) significant minder hoog scoren dan de scholen met minder achterstands-leerlingen. Het effect is klein. Het verschil tussen weinig tot geen achterstandsleerlingen (<10%; stratum 1) en de groep met wat meer gewichtenleerlingen (10-25%; stratum 2) is hier niet significant. Een uitzondering is de vaardigheid Spelling: daar doen de stratum 2-scholen het beter dan de stratum 1-scholen. Merk wel op dat dit gecorrigeerde effecten zijn. Het gaat om het effect van school bovenop de effecten van de individuele gewichtenleerlingen. Als daar niet voor gecorrigeerd wordt is het effect bij Spelling kleiner. De verschillen bij de andere vaardigheden worden dan juist over het algemeen groter. Als we de effecten vergelijken met de effecten zoals die in 2011 gevonden zijn (zie Hemker en Van Weerden, 2012) dan zien we dat de resultaten voor 2012 en 2011 erg op elkaar lijken. Nuanceverschillen zijn te vinden bij Spelling en Rekenen, waar in 2011 de stratum 3-scholen relatief iets beter presteerden dan in 2012.

Regio

Het contrast tussen de onderscheiden regio's levert slechts incidenteel een significant verschil op. Dat is alleen het geval bij Spelling voor het contrast Oost- West. De effectgrootte is echter niet betekenisvol (>.20).

Verstedelijking

Voor de variabele verstedelijking zijn er geen contrasten die een significant verschil opleveren (p=0,05).

Tabel 3.7 Effectgrootten van achtergrondvariabelen op schoolniveau voor 2012**

Variabele	Contrast	Rekenen-Wiskunde	Woordenschat	Begrijpend lezen	Spelling
Stratum	Stratum 1-Stratum 2	0,05	-0,04	0,01	**-0,15**
	Stratum 1-Stratum 3	**0,31**	**0,22**	**0,33**	-0,07
	Stratum 2-Sratum 3	**0,26**	**0,25**	**0,32**	0,08
Regio	Noord – Oost	0,15	0,14	0,05	-0,04
	Noord – West	0,17	0,16	0,08	-0,17
	Noord – Zuid	0,18	0,18	0,10	-0,03
	Oost – West	0,01	0,01	0,03	-0,13
	Oost – Zuid	0,02	0,03	0,05	0,02
	West –Zuid	0,01	0,02	0,02	0,14
Verstedelijking	stad-land	-0,07	-0,04	-0,03	0,06

* vet= significant (p=0.01); cursief =significant (p=0.05)

4 Conclusies

4.1 Algemeen beeld

Vergelijken we de prestaties van 2012 met die van 2011, dan blijken de prestaties van de leerlingen in jaargroep 8 overwegend te zijn verbeterd. Bij taal is dat voor twee van de drie vaardigheden duidelijk het geval, nl. bij Spelling en in mindere mate bij Woordenschat. Bij Begrijpend lezen is het beeld gecompliceerd. Vergelijken we met behulp van de ankertoets dan lijkt er sprake van een daling. Nemen we daarentegen de overeenkomstige items van de toets Basisvaardigheden als basis, dan lijken de prestaties te stijgen. De vergelijking met de ankertoets is de meest conservatieve en is het best vergelijkbaar met wat in eerdere peilingen is uitgevoerd. Daarmee is het niveau van leesvaardigheid vrijwel gelijk aan dat van 2010. Er is dan wel sprake van een stijging ten opzichte van 2008. De volgende meting moet uitwijzen welke kant de trend op gaat.
Voor de rekenonderwerpen is de situatie consistent en eenduidig positief.

De hele periode van vijf jaar overziend blijkt dat bij alle onderdelen er ten opzichte van de start van de meting in 2008 vooruitgang is te bespeuren. Dat is bij het ene onderwerp wat meer (Spelling, de rekenvaardigheden) dan bij het andere (Begrijpend lezen, Woordenschat), maar de trend is overal positief; bij alle meetschalen is de waarde nu gemiddeld hoger dan het uitgangsgetal van 250. Dat is ook nog steeds zo bij Begrijpend lezen, ook als we uitgaan van de conservatieve aanpak en interpretatie. Wel moet geconstateerd worden dat voor alle vaardigheden in jaargroep 8 de effectgroottes van de reguliere vergelijkingen nog steeds onder de grens van 0.20 vallen en dus niet betekenisvol mogen worden genoemd.

Bij jaargroep 4 zien we een vergelijkbaar beeld als bij jaargroep 8. Positieve effecten zien we vooral bij Woordenschat, Spelling en Rekenen-Wiskunde. Begrijpend lezen is hier een vaardigheid waar relatief weinig vooruitgang gevonden is: de prestaties lijken weer iets te dalen, maar blijven nog wel duidelijk boven het startniveau van 2008. Bij alle onderwerpen is de trend ten opzichte van het startjaar 2008 dus positief, bij alle meetschalen zijn de waarden nu hoger dan de 250 van het beginjaar.

De effectgroottes zijn nu in jaargroep 4, anders dan in 2011, groot genoeg om betekenisvol genoemd te mogen worden, als we daarbij de kwalificaties volgen die in de evaluatieliteratuur gebruikelijk zijn. De drempel van 0.20 die staat voor een klein effect, wordt in drie van de vier gevallen gehaald. Alleen bij Begrijpend lezen is dat dus niet het geval. Het verschil met 2008 is zelfs niet significant.

Nemen we de twee jaargroepen bij elkaar dan zien we dat er voor Rekenen-Wiskunde sprake is van een eenduidige positieve trend. Die is in jaargroep 4 groter dan in jaargroep 8, maar in beide gevallen onmiskenbaar.
Voor Spelling is eveneens in beide gevallen een eenduidige positieve ontwikkeling te melden die ook in jaargroep 4 wat sterker is dan in jaargroep 8. Woordenschat laat in jaargroep 4 eveneens een duidelijke stijging zien met dezelfde omvang als Spelling en Rekenen-Wiskunde, maar vertoont in jaargroep 8 een wat grilliger verloop, met uiteindelijk zeer kleine, zij het wel significante, vooruitgang ten opzichte van 2008.

Voor Begrijpend lezen is de situatie nog lastiger te duiden en wellicht zorgelijk. In jaargroep 4 is er nauwelijks progressie te zien vergeleken met 2008, nu er vanaf 2011 weer een lichte daling is geconstateerd. In jaargroep 8 is de situatie afhankelijk van de vergelijkingsmethode. De meest conservatieve schatting laat echter nog steeds een (zeer) kleine vooruitgang zien ten opzichte van 2008.

Alles bij elkaar genomen mogen we dus concluderen dat de trend in alle gevallen positief uitvalt, waarbij we de grootse vooruitgang aantreffen in jaargroep 4. Alleen de ontwikkeling bij Begrijpend lezen is diffuus en moeilijk te duiden.

4.2 Invloed van achtergrondvariabelen

Jongens in jaargroep 8 zijn beter in Rekenen-Wiskunde dan meisjes. Meisjes zijn beter in Spelling.
Bij Begrijpend lezen en Woordenschat zien we geen opvallend verschil. In jaargroep 4 zien we dezelfde verschillen. Zowel Woordenschat als Begrijpend lezen leveren bij meisjes een hogere score op, een effect dat we in jaargroep 8 niet meer terugzien.
Vertraagde leerlingen blijken in alle gevallen een grote achterstand te hebben. De effectgrootte is bij jaargroep 8 forser dan in jaargroep 4 en over het algemeen spreken we van een klein tot matig effect.
In vergelijking met de andere onderwerpen zien we bij Woordenschat het kleinste verschil.

In jaargroep 8 zien we dat leerlingen zonder formatiegewicht (0.00) het op bijna alle vaardigheden beter doen dan leerlingen mét een formatiegewicht, laag (0.30) of hoog (1.20). Het gaat dan om kleine en matige effectgroottes. Alleen bij Spelling is dat verschil niet significant. Het blijkt dat hier ook na correctie leerlingen met een hoog formatiegewicht (1.20) significant hoger scoren dan leerlingen met een laag formatiegewicht, een verschil in onverwachte richting. Woordenschat is de enige vaardigheid waarbij leerlingen met een laag formatiegewicht significant hoger scoren dan leerlingen met een hoog formatiegewicht. Er is daar sprake van een klein effect.
In jaargroep 4 is het beeld hetzelfde, maar is het contrast tussen de leerlingen met hoog of laag formatiegewicht bijna nergens significant, alleen bij Woordenschat (p=0,05). Ook hier vertoont Spelling eenzelfde afwezigheid van effect bij het contrast tussen hoog en laag formatiegewicht.

De contrasten voor stratum leveren in jaargroep 8 en jaargroep 4 een min of meer vergelijkbaar beeld op, maar met kleine verschillen. Bij Woordenschat en Begrijpend lezen worden in beide leerjaren kleine significante verschillen gevonden tussen de stratum 3-scholen (veel gewichtsleerlingen) en de stratum 1- en stratum 2-scholen. De scholen met veel gewichtsleerlingen presteren slechter. Bij Spelling waren deze effecten in beide leerjaren niet significant. Bij de rekenvaardigheden was in jaargroep 4 het resultaat vergelijkbaar met Woordenschat en Begrijpend lezen. In jaargroep 8 het verschil tussen weinig en veel gewichtsleerlingen op een school significant voor 2 van de 3 rekenvaardigheden. Alle overige verschillen tussen scholen waren niet significant bij Rekenen.
De verschillen tussen stratum-1 en stratum-2 scholen was in beide jaargroepen bij bijna alle vaardigheden te klein om tot een betekenisvol effect te komen. Enige uitzondering was de vaardigheid Spelling in jaargroep 4. Daar presteerden de stratum-2 scholen beter.

De schoolvariabelen regio en verstedelijking leveren alleen in jaargroep 4 een incidenteel effect op en wel bij Rekenen-Wiskunde.

Een variabele in jaargroep 8 die wel veel significante verschillen oplevert en ook grote effecten laat zien is het doorstroomadvies voor het VO. De grootste waarden zien we bij Begrijpend lezen en Rekenen-Wiskunde, waar het contrast tussen leerlingen met vwo-advies en leerlingen met vmbo BB-advies bijna vijf standaarddeviaties beslaat. Dat zijn in de context van de sociale wetenschappen en met name die van onderwijsonderzoek enorme verschillen. Het geeft aan hoe groot de vaardigheidsverschillen aan het eind van de basisschool al zijn. Voor Spelling is het verschil dan nog het kleinst, nl. 3,4 standaarddeviaties. Vooral het verschil tussen de leerlingen met adviezen vwo en havo levert een grote bijdrage aan de grootte van het totaaleffect.

Interessant is het verschil in groei over de afgelopen vijf jaar bij leerlingen met verschillende achtergronden. Het verschil tussen jongens en meisjes lijkt vrij constant, zeker in jaargroep 4. Opvallend is wel dat de verschillen in omvang in termen van vaardigheidsscores groter blijken te zijn in jaargroep 4 dan in jaargroep 8. Het lijkt er op dat het verschil ten gevolge van geslacht in de loop van de basisschool, in ieder geval van jaargroep 4 naar jaargroep 8, dus afneemt.
Bij de categorieën voor verschil in de gebruikte thuistaal is de ontwikkeling minder constant. Anders dan in de rapportage van vorig jaar, lopen de leerlingen die thuis tweetalig zijn hun achterstand niet verder in. Alleen bij deze groep lijkt de score weer te zakken. Bij leerlingen die thuis alleen buitenlands spreken is dat niet het geval.

Dat is het meest zichtbaar bij Rekenen-Wiskunde en Woordenschat, maar we zien het ook bij Begrijpend lezen en bij Spelling. In jaargroep 8 kan dat voor 2012 niet worden gevolgd omdat deze variabele in de achtergrondvariabelen voor de Eindtoets van dat jaar ontbrak.

4.3 Discussie

Alles bij elkaar genomen mogen we concluderen dat de trend voor alle reken- en taalvaardigheden positief uitvalt zowel in jaargroep 4 als 8, waarbij we de grootse vooruitgang aantreffen in jaargroep 4.
De ontwikkelingen bij Begrijpend lezen en Woordenschat roepen echter vraagtekens op. Ook valt op dat leerlingen met als thuistaal een buitenlandse taal hun achterstand niet verder lijken in te lopen.

Opvallend is ook de uitkomst bij Spelling. De stijging van de score op Spelling is in jaargroep 4 het grootst van alle vaardigheden. Als we de achtergrondvariabelen daarbij betrekken, dan valt op dat, anders dan bij de andere vaardigheden, thuistaal daar nauwelijks invloed op heeft. Ook de variabele herkomst blijkt er weinig toe te doen. Wel lijkt het dat jongens bij Spelling meer progressie vertonen dan meisjes, vooral in jaargroep 4. Afwijkend is ook de uitkomst gedifferentieerd naar formatiegewicht. Leerlingen met een hoog formatiegewicht (1.2) hebben bij Spelling een hogere vaardigheid dan leerlingen met een laag formatiegewicht (0.3), terwijl het verschil met leerlingen zonder formatiegewicht niet significant is. Bij de andere reken- en taalvaardigheden geeft dat laatste contrast juist het grootste effect te zien.

De resultaten zijn nogal verschillend voor de beide jaargroepen. De vooruitgang in jaargroep 4 is groter en heeft een in het algemeen een minder grillig verloop dan in jaargroep 8. Wellicht zijn hier verklaringen voor te vinden, maar die kunnen nu nog niet gefundeerd gegeven worden. Het verschil in uitkomsten bij jaargroep 8 en jaargroep 4 onderstreept wel het belang van een monitoring voor beide jaargroepen afzonderlijk, zoals tot nu toe is gebeurd.

Opvallend is de min of meer lineaire vooruitgang bij Rekenen-Wiskunde en Spelling over de afgelopen vijf jaar. De vaardigheden Woordenschat en Begrijpend lezen vertonen daarentegen een ontwikkeling die er grillig en diffuus uitziet. Wellicht dat dit verschil te maken heeft met het verschil in type vaardigheid, waarbij de aard van het gemeten construct en de mate van beïnvloedbaarheid door de school een rol zou kunnen spelen, maar dat zou verder onderzoek moeten uitwijzen. Ook is het bij Woordenschat voorstelbaar dat de meting van dit construct in de Eindtoets minder robuust en stabiel is dan wenselijk voor een afzonderlijke rapportage. De toets bestaat daar uit 20 items, terwijl de LVS-toets uit 60 items bestaat. Dat heeft ook zijn weerslag op de psychometrische kwaliteit.[4]

Hoewel deze evaluatie gestart is in het kader van de Kwaliteitsagenda (2008), waarin o.a. het opbrengstgericht werken wordt gestimuleerd, met name bij taal en rekenen, kan op grond van de uitkomsten van dit onderzoek slechts ten dele een adequate beleidsevaluatie worden gepleegd. Voor het vinden van verklaringen ontbreken gegevens over wat scholen precies hebben gedaan op het gebied van de gemeten vaardigheden. We kunnen dus niet vaststellen in hoeverre hierin wijzigingen zijn opgetreden in de loop der jaren.
Het verdient daarom aanbeveling om, net als bij het regulier peilingsonderzoek (PPON) gebruikelijk is, het onderwijsaanbod of veranderingen daarin te inventariseren zodanig dat gegevens betreffende de leerlingprestaties daaraan gekoppeld kunnen worden. Met behulp van informatie over bijvoorbeeld het gebruik van andere methoden of het besteden van meer tijd aan bepaalde onderwerpen kan wellicht een plausibele interpretatie worden verbonden aan de verandering in resultaten bij leerlingen.

[4] Het verschil tussen deze toetsen is na te lezen in de inhoudelijke beschrijving daarvan in de rapportage van 2008 (Hemker & Van Weerden, 2009).

Literatuur

Berkel, S. van, M. Hilte, R. Engelen, F. Kamphuis, F. Kleintjes, R. Krom (2010). *Woordenschat Groep 3 t/m 5. Wetenschappelijke verantwoording.* Cito, Arnhem.

Boxtel, H. van, R. Engelen, A. de Wijs (2012). *Wetenschappelijke verantwoording van de Eindtoets Basisonderwijs 2010*. Cito, Arnhem.

Cito (2012). *Terugblik en resultaten 2012. Eindtoets basisonderwijs jaargroep 8*. Cito, Arnhem.

Cohen, J. (1988): *Statistical Power Analysis for the Behavioral Sciences (second ed.)*. Lawrence Erlbaum Associates.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008). *Over de drempels met taal en rekenen.* SLO, Enschede.

Feenstra, H., F. Kamphuis, F. Kleintjes, R. Krom (2010). *Wetenschappelijke verantwoording Begrijpend lezen voor groep 3 tot en met 6.* Cito, Arnhem.

Hemker, B.T. (2012). *The impact of motivation: modeling motivation in educational measurement. Presentation presented July 4, 2012 at the ITC conference, Amsterdam.*

Hemker, B.T., J. Kordes & J.J. van Weerden (2011). *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2010 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau*. Cito, Arnhem. (zie www.cito.nl, Onderzoek en wetenschap, PPON)

Hemker, B.T., J.B. Kuhlemeier & J.J. van Weerden (2010). *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2009 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau*. Cito, Arnhem. (zie www.cito.nl, Onderzoek en wetenschap, PPON)

Hemker, B.T. & J.J. van Weerden (2009). *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2008 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau - Technische rapportage*. Cito, Arnhem. (http://www.minocw.nl/documenten/133682d.pdf)

Hemker, B.T. & J.J. van Weerden (2012). *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2011 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau - Technische rapportage*. Cito, Arnhem. (zie www.cito.nl, Onderzoek en wetenschap, PPON)

Inspectie van het onderwijs (2012). *Monitor verbetertrajecten taal en rekenen 2008/2009, 2009/2010 en 2010/2011,* Inspectie van het Onderwijs/OCW, Utrecht.

Janssen, Jan, Frank van der Schoot, Bas Hemker (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 4. Uitkomst van de vierde peiling in 2004*. PPON-reeks nummer 32. Cito, Arnhem. (http://www.cito.nl/po/ppon/rekwisk/eind_fr.htm)

Janssen, J., N. Verhelst, R. Engelen en F. Scheltens (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8.* Cito, Arnhem.

Verhelst, N.D. (1993). Itemresponstheorie. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk.* Arnhem: Cito (p. 83-178).

Verhelst, N.D., C.A.W. Glas & H.H.F.M. Verstralen (1995). *OPLM: One Parameter Logistic Model. Computer program and manual.* Arnhem: CITO.

Verhelst, N.D. & H.H.F.M. Verstralen (2002). *Structural Analysis of a Univariate Latent Variable (SAUL); Theory and a Computer Program.* Arnhem: Cito.

Wijs, A. de, F. Kamphuis, F. Kleintjes, M. Tomesen (2010). *Wetenschappelijke verantwoording Spelling voor groep 3 tot en met 6.* Cito, Arnhem.

Relevante websites:

www.cito.nl
www.toetswijzer.nl
http://ppon.cito.nl
www.minocw.nl

Bijlagen

Bijlage 1 Gemiddelden en standaarddeviaties per vaardigheid gecategoriseerd naar achtergrondvariabele in jaargroep 8 in 2012

	Vaardigheid*	Taal								Rekenen-Wiskunde					
		Sp		BL-A		BL-B		WS		GB		BPV		MMTG	
Variabele:	**Categorie:**	**Gem.**	**SD**	**Gem.**	**SD**	**Gem.**	**SD**	**Gem.**	**SD**	**Gem.**	**SD**	**Gem.**	**SD**	**Gem.**	**SD**
Jaar		**256**	**50**	**253**	**50**	**262**	**50**	252	50	**255**	**50**	**257**	**50**	**256**	**50**
Geslacht	Jongens	249	49	250	50	259	50	252	50	265	49	266	48	265	49
	Meisjes	262	50	256	50	265	50	252	50	245	49	248	48	248	49
Leertijd	Regulier	262	48	259	48	268	48	257	49	261	48	263	47	262	48
	Vertraagd	227	48	225	49	233	49	228	50	228	49	228	48	226	49
Stratum	Stratum (lovs*) 1	259	50	261	48	270	48	259	48	259	49	263	48	262	49
	Stratum (lovs*) 2	253	50	249	49	258	49	250	49	251	50	253	49	252	50
	Stratum (lovs*) 3	247	51	226	50	235	50	225	51	243	51	240	50	237	51
	F 0.00	258	49	259	48	268	48	257	48	259	49	261	48	261	48
	F 0.30	237	50	222	49	230	49	227	49	229	50	228	48	227	49
	F 1.20	246	51	215	49	223	49	206	49	336	50	232	49	227	49
Soort toets	EB	257	49	255	48	265	48	254	49	258	47	259	47	259	47
	NT	208	50	175	50	176	49	187	51	161	48	173	48	167	49
Advies VO	vmbo-BB	199	41	175	37	182	37	185	41	179	38	180	36	179	36
	vmbo-KB	218	41	209	37	218	37	214	40	214	38	213	36	212	36
	vmbo-GT	232	40	231	37	240	37	231	40	234	38	235	36	232	36
	havo	265	40	272	36	280	36	264	40	271	37	274	36	272	35
	vwo	318	40	317	36	325	36	312	39	316	37	322	36	325	35
Regio	Noord	253	50	250	49	259	49	256	49	252	50	254	50	254	49
	Oost	253	50	254	49	263	49	253	49	254	50	257	50	256	49
	West	257	50	252	51	261	51	251	51	256	50	257	50	256	49
	Zuid	258	50	255	50	264	50	251	50	257	50	258	50	257	49
Urbanisatie-	Zeer Sterk	257	51	244	52	252	52	243	53	252	51	252	51	250	51
graad	Sterk	256	50	253	50	262	50	251	50	256	50	257	50	256	50
	Matig	258	50	256	49	265	49	254	49	257	50	259	50	259	50
	Weinig	255	50	255	49	265	49	254	49	255	50	258	49	257	49
	Niet	253	50	254	49	263	49	256	49	253	50	256	49	257	49
IJK	Langer dan 4 jr in NL	260	48	258	47	267	47	256	48	260	47	262	47	261	47
	Korter dan 4 jr in NL	213	49	193	49	200	48	201	50	194	48	197	48	197	48
Schoolgrootte	Klein (<200 lln)	251	50	249	50	258	50	251	50	251	50	253	49	253	50
	Groot (200+ lln)	258	50	255	50	264	50	252	50	257	50	259	49	257	50

* Sp = Spelling; BL-A = Begrijpend lezen (anker); BL-B = Begrijpend lezen (toets BV); WS = Woordenschat; G/B = Getallen en bewerkingen; BPV = Breuken, procenten, verhoudingen; MMTG = Meetkunde, meten, tijd en geld.

Bijlage 2 Effectschattingen voor alle vaardigheden gecategoriseerd naar achtergrondvariabelen in jaargroep 8 in 2012

Variabele	Effecten 2012	Sp		BL-A		BL-B		WS		G/B		BPV		MMTG	
	Contrast	Eff.gr	sign	Eff.gr	sign	Eff.gr	sign	Eff.gr	sign	Eff.gr	sign	Eff.gr	sign	Eff.gr	sign
Geslacht	Meisjes - Jongens	0,25	***	0,12	***	0,11	***	0,00		-0,43	***	-0,43	***	-0,39	***
Leertijd	Vertraagd - Regulier	-0,69	***	-0,60		-0,61	***	-0,50	***	-0,67	***	-0,71	***	-0,71	***
Stratum	Stratum: Nieuw 2 – Nieuw 1	-0,07	***	-0,17	***	-0,17	***	-0,12	***	-0,09	***	-0,12	***	-0,12	***
	Stratum: Nieuw 3 – Nieuw 2	-0,05		-0,29	***	-0,29	***	-0,28	***	-0,04		-0,11	***	-0,15	***
	Stratum: Nieuw 3 – Nieuw 1	-0,12	***	-0,45	***	-0,45	***	-0,40	***	-0,14	***	-0,24	***	-0,27	***
Formatie-gewicht	F 0.30 – F 0.00	-0,33	***	-0,59	v	-0,6	***	-0,48	***	-0,49	***	-0,56	***	-0,56	***
	F 1.20 – F 0.30	0,27	***	0,03		0,02		-0,28	***	0,24	***	0,20	***	0,12	*
	F 1.20 – F 0.00	-0,06		-0,57	***	-0,58	***	-0,76	***	-0,27	***	-0,36	***	-0,43	***
Soort toets	NT - EB	-0,79	***	**-1,45**	***	**-1,61**	***	**-1,10**	***	**-1,86**	***	**-1,62**	***	**-1,71**	***
Advies VO	vmbo-KB - vmbo-BB	0,44	***	**0,85**	***	**0,91**	***	0,65	***	**0,93**	***	**0,89**	***	**1,13**	***
	vmbo-GT - vmbo-KB	0,34	***	0,54	***	0,46		0,37	***	**0,86**	***	0,55	***	0,67	***
	havo - vmbo-GT	0,78	***	**1,07**	***	**1,06**	***	0,79	***	**0,98**	***	**1,07**	***	**1,36**	***
	vwo - havo	**1,36**	***	**1,23**	***	**1,21**	***	**1,17**	***	**1,20**	***	**1,33**	***	**1,83**	***
Regio	Oost - Noord	-0,01		0,10	**	0,10	**	-0,05		0,04		0,06		0,06	
	West - Noord	0,06		0,13	***	0,13	***	0,00		0,09	**	0,09	**	0,09	**
	Zuid - Noord	0,11	***	0,14	***	0,13	***	-0,07		0,11	***	0,09	**	0,10	**
	West - Oost	0,07	**	0,03		0,03		0,05	*	0,05	*	0,02		0,03	
	Zuid - Oost	0,12	***	0,04		0,04		-0,02		0,06	**	0,03		0,04	
	Zuid - West	0,05		0,01		0,00		-0,07	**	0,01		0,01		0,01	
Urbanisatie-graad	Sterk - Zeer sterk	-0,11	***	-0,01		-0,01		-0,01		-0,03		-0,04		-0,04	
	Matig - Sterk	0,01		0,02		0,02		0,00		0,01		0,01		0,03	
	Weinig - Matig	-0,05		-0,02		-0,02		-0,01		-0,03		-0,02		-0,05	
	Niet - Weinig	-0,06		-0,04		-0,04		0,00		-0,05		-0,05		0,01	
IJK	Wel - geen IJK code	**-0,81**	***	**-1,15**	***	**-1,21**	***	**-0,92**	***	**-1,37**	***	**-1,22**	***	**-1,16**	***
Schoolgrootte	Groot - Klein	0,13	***	0,07	***	0,07	***	-0,03		0,08	***	0,07	***	0,05	*

- alle effecten groter dan ,20 of kleiner dan -,20 zijn significant met α < 0,0001 (niet meer afzonderlijk aan gegeven)
- geclassificeerd als “geen effect”, maar wel significant:* : 0,01 < α < -0,001; **: 0,001 < α < 0,0001; ***: α < 0,0001
- Sp = Spelling; BL-A = Begrijpend lezen (anker); BL-B = Begrijpend lezen (toets BV); WS = Woordenschat; G/B = Getallen en bewerkingen; BPV = Breuken, procenten, verhoudingen; MMTG = Meetkunde, meten, tijd en geld.

Bijlage 3 Gemiddelden en standaarddeviaties per vaardigheid gecategoriseerd naar achtergrondvariabele in jaargroep 4 in 2012

Groep 4	jaar 2012		Rekenen		Begrijpend lezen		Spelling		Woordenschat	
	n	%	Gem.	SD	Gem.	SD	Gem.	SD	Gem.	SD
Totaal	2394	100	261	52	252	47	262	45	261	48
Naar achtergrondvariabele:										
Geslacht										
Jongen	1218	51%	270	54	247	48	257	47	258	49
Meisje	1176	49%	253	47	257	45	267	43	264	48
Leertijd*										
Regulier	2093	87%	264	51	254	47	264	46	263	48
Vertraagd	301	13%	245	50	235	46	246	40	247	48
Formatiegewicht										
0.00	2125	89%	265	51	255	47	263	46	264	48
0.30	146	6%	238	49	233	43	250	37	244	43
1.20	123	5%	236	48	226	42	257	44	233	47
Herkomst										
Nederland	2222	93%	263	51	254	47	261	45	263	48
Niet-Nederlands	172	7%	238	51	230	44	269	45	237	48
Stratum										
Stratum 1	1569	66%	266	51	255	46	261	45	264	47
Stratum 2	490	20%	261	51	253	48	267	47	264	50
Stratum 3	335	14%	243	53	234	43	259	45	246	49
Regio										
Noord	146	6%	270	48	256	41	258	41	270	48
Oost	294	12%	263	52	255	47	259	43	264	50
West	1533	64%	260	52	251	46	263	46	260	48
Zuid	421	18%	262	52	253	50	259	45	262	48
Verstedelijking										
Zeer sterk	284	12%	253	59	246	49	269	52	263	56
Sterk	921	38%	261	51	251	47	260	44	258	47
Matig	408	17%	261	48	251	47	257	44	261	46
Weinig	516	22%	264	50	256	47	263	46	264	50
Niet	265	11%	268	52	256	44	266	46	266	43
Stad [(zeer)sterk]	1205	50%	259	53	250	48	262	46	259	50
Land [matig-niet]	1189	50%	264	50	254	46	261	45	263	47
Thuistaal										
Nederlands	2182	91%	264	51	255	47	262	45	265	47
Mengeling	121	5%	230	48	226	41	255	45	230	44
Buitenlands	91	4%	232	45	221	38	257	43	217	39

* afhankelijk van groep

Bijlage 4 Effectschattingen voor alle vaardigheden gecategoriseerd naar achtergrondvariabelen in jaargroep 4 in 2012 (gecorrigeerd)*

		Rekenen-Wiskunde		Woordenschat		Begrijpend lezen		Spelling	
Variabele	**Contrast**	**Effect**	**Sign**	**Effect**	**Sign**	**Effect**	**Sign**	**Effect**	**Sign**
Geslacht	Jongens-meisjes	0,36	***	-0,10	*	-0,21	***	-0,20	***
Leertijd	Regulier-Vertraagd	0,34	***	0,24	***	0,30	***	0,36	***
Formatiegewicht	0.0 - 0.3	0,41	***	0,34	***	0,32	***	0,28	**
	0.0 - 1.2	0,40	***	0,51	***	0,41	***	0,06	
	0.3 - 1.2	0,00		0,18		0,09		-0,22	
Herkomst	Nederland-Niet-Neder	0,24	*	0,30	**	0,23	*	-0,32	**
Thuistaal	Alleen Nls - Nls & Bui	0,54	***	0,69	***	0,44	***	0,18	***
	Alleen Nls - Alleen Bui	0,45	***	0,90	***	0,51	***	0,09	***
	NLs & Bui - Alleen Bui	-0,09		0,21		0,07		-0,09	
Stratum	Stratum 1-Stratum 2	0,05		-0,04		0,01		-0,17	**
	Stratum 1-Stratum 3	0,31	***	0,22	**	0,33	***	-0,04	
	Stratum 2-Stratum 3	0,26	***	0,25	**	0,32	***	0,13	
Regio	Noord – Oost	0,15		0,14		0,05		0,01	
	Noord – West	0,17		0,16		0,08		-0,14	
	Noord – Zuid	0,18		0,18		0,10		-0,02	
	Oost – West	0,01		0,01		0,03		-0,15	*
	Oost – Zuid	0,02		0,03		0,05		-0,03	
	West –Zuid	0,01		0,02		0,02		0,11	*
Verstedelijking	stad-land	-0,07		-0,04		-0,03		0,05	

* Indien contrast niet significant, dan effectgrootte niet relevant.

significantie	
p =0,05 tot 0,01	*
p=0,01 tot 0,001	**
p <=0,001	***

effectgrootte	
	klein
	matig
	groot

Cito maakt wereldwijd werk van goed en eerlijk toetsen en beoordelen. Met de meet- en volgmethoden van Cito krijgen mensen een objectief beeld van kennis, vaardigheden en competenties. Hierdoor zijn verantwoorde keuzes op het gebied van persoonlijke en professionele ontwikkeling mogelijk. Onze expertise zetten we niet alleen in voor ons eigen werk maar ook om advies, ondersteuning en onderzoek te bieden aan anderen.

Cito
Amsterdamseweg 13
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
F (026) 352 13 56
www.cito.nl

Klantenservice
T (026) 352 11 11
klantenservice@cito.nl

Fotografie: Ron Steemers