

Wetenschappelijke verantwoording van de Eindtoets Basisonderwijs 2010

Herman van Boxtel, Ronald Engelen, Anja de Wijs



Wetenschappelijke verantwoording van de Eindtoets 2010

Herman van Boxtel
Ronald Engelen
Anja de Wijs

© Cito B.V. Arnhem (2011)

Niets uit dit werk mag zonder voorafgaande schriftelijke toestemming van Cito worden openbaar gemaakt en/of verveelvoudigd door middel van druk, fotokopie, scanning, computersoftware of andere elektronische verveelvoudiging of openbaarmaking, microfilm, geluidskopie, film- of videokopie of op welke wijze dan ook.

Inhoud

1	Inleiding	5
2	Uitgangspunten van de toetsconstructie	9
2.1	Meetpretentie	9
2.2	Doelgroep	9
2.3	Gebruiksdoel en functie	9
2.4	Theoretische inkadering	10
2.4.1	Inhoudelijk	10
2.4.2	Psychometrisch	12
2.4.2.1	Constructie- en equivaleringsprocedures	12
2.4.2.2	Het gehanteerde meetmodel	19
3	Beschrijving van de toets	23
3.1	Opbouw en structuur	23
3.2	Inhoudsverantwoording	24
3.2.1	Taal	25
3.2.2	Rekenen-Wiskunde	28
3.2.3	Studievaardigheden	31
3.2.4	Wereldoriëntatie	34
3.3	Kenmerken van de toets: beschrijvende gegevens	38
4	Het normeringsonderzoek	43
4.1	De Eindtoets: het specifieke karakter van de normering	43
4.2	Kenmerken van Eindtoetsgebruikers in vergelijking met de gehele doelpopulatie	44
4.3	Kalibratie en normering	48
4.3.1	De stappen in de kalibratie	48
4.3.2	Evaluatie van de kalibratieprocedure	49
4.3.3	Normering en adviescategorieën	54
5	Betrouwbaarheid en meetnauwkeurigheid	57
5.1	Betrouwbaarheid	57
5.2	Nauwkeurigheid	59
6	Validiteit	63
6.1	Inhoudsvaliditeit	63
6.2	Begripsvaliditeit	64
6.2.1	Passing van het meetmodel; psychometrische kwaliteit van de opgaven	64
6.2.2	Structuur	64
6.2.3	Equivalentie met eerdere toetsversies	66
6.2.4	Soortgenootvaliditeit; convergente en discriminante validiteit	67
6.2.5	De Eindtoets over de jaren heen en in relatie tot diverse subgroepen	69
6.2.6	De Eindtoets en allochtone leerlingen; DIF-analyses	78
6.3	Criteriumvaliditeit	80
6.3.1	Relatie met het doorstroomadvies van de leerkracht	81
6.3.2	Het doorstroomadvies: intelligentie en leervorderingen vergeleken	82
6.3.3	De score op de Eindtoets en de feitelijke plaatsing en doorstroom	86
7	Samenvatting	91
8	Literatuur	95

1 Inleiding

Sinds 1970 brengt Cito de Eindtoets Basisonderwijs uit, als opvolger van de Amsterdamse Schooltoets. Elk jaar wordt sindsdien een volledig nieuwe versie geconstrueerd en afgenomen bij leerlingen van groep 8 van het reguliere basisonderwijs. Dat heeft niet elk jaar geleid tot een nieuwe wetenschappelijke verantwoording. In de jaren zeventig en tachtig verschenen dergelijke verantwoordingen met enige regelmaat (een interessante is bijvoorbeeld die voor de jaargang 1981 vanwege het historisch overzicht over de periode 1966-1980; zie Wijnstra, 1984). De laatst verschenen verantwoording is die over het jaar 1992 (Uiterwijk & Theunissen, 1996), maar de laatste door de COTAN beoordeelde versie van de Eindtoets betreft de jaargang 1987 op basis van een beoordeling uit 1992 (zie Evers, A., Braak, M.S.L., Frima, R.M., & Vliet-Mulder, J.C. van (2009)). Een jaarlijks vernieuwde versie van de wetenschappelijke verantwoording is ook eigenlijk niet echt nodig, enerzijds omdat de vernieuwing van de inhoud steeds plaatsvindt op basis van een gedetailleerde catalogus van onderwijsdoelstellingen en geavanceerde schalingsmethoden, anderzijds omdat de toets genormeerd wordt op zeer omvangrijke groepen leerlingen die voor een groot deel samenvallen met de populatie zelf. Deze toetskenmerken zorgen ervoor dat resultaten van validerings- en ander onderzoek dat is uitgevoerd met eerdere versies van het instrument, generaliseerbaar zijn naar elke nieuwe (equivalente) toetsversie.

Niettemin is er voldoende aanleiding om bij de Eindtoets Basisonderwijs 2010 deze nieuwe wetenschappelijke verantwoording uit te brengen. Argumenten daarvoor zijn als volgt samen te vatten:

- Op de eerste plaats zijn er in de loop van de tijd kleine veranderingen aangebracht in de structuur en opzet van de toets. Daarmee is de toets niet meer volledig identiek aan de versie van 1970. Een overzicht van de wijzigingen en een toelichting op het waarom daarvan maakt daarom deel uit van deze verantwoording.
- De (ontkennende) beantwoording van de vraag of er jaarlijks een nieuwe verantwoording nodig is, hangt af van de deugdelijkheid van bovengenoemde procedures en technieken. Deze zijn immers bedoeld om te leiden tot volstrekt equivalente versies van de toets. Cito dient deze procedures en technieken zo volledig mogelijk te beschrijven, ook als daarin wijzigingen worden aangebracht. Pas dan kan de gebruiker nagaan of de pretenties van de toetsconstructeurs in termen van equivalentie gerechtvaardigd zijn. Deze verantwoording bevat een dergelijke beschrijving.
- Een beschrijving van de procedures volstaat echter niet. Men zal ook empirisch moeten onderbouwen dat de procedures en technieken leiden tot verregaand equivalente versies. Deze onderbouwing is inmiddels beschikbaar (Staphorsius, Béguin, Engelen, Kleintjes & Verhelst, 2007) en wordt in deze verantwoording gerapporteerd. Als zodanig vormt de empirische onderbouwing van equivalentie op zichzelf al een bijdrage aan de inhouds- en begripsvaliditeit van de toets.
- In de loop der jaren is er veel onderzoek gedaan naar en met de Eindtoets Basisonderwijs op het punt van begrips- en criteriumvaliditeit, niet alleen op het punt van de equivalering. Zo zijn de uitkomsten van de toets bijvoorbeeld gekoppeld aan die van een eerder afgenomen Entreetoets, aan intelligentiescores van de leerlingen en aan het doorstroomadvies van de leerkracht. Ook is follow-up onderzoek gedaan in termen van feitelijke doorstroomgegevens. Verder zijn analyses uitgevoerd op het effect van factoren als bijvoorbeeld gezinsinkomen, herkomst en thuistaal. Veel analyses worden jaarlijks gerapporteerd in de zogenoemde 'Terugblik'. Juist vanwege de equivalentie van de verschillende toetsversies accumuleert het empirisch bewijsmateriaal. Deze wetenschappelijke verantwoording probeert de resultaten die in de afgelopen jaren beschikbaar zijn gekomen op een overzichtelijke manier bij elkaar te zetten, dit naast de gegevens die specifiek beschikbaar zijn voor de Eindtoets Basisonderwijs versie 2010.

Deze verantwoording bevat alle informatie die nodig is voor een beoordeling van de kwaliteit van de Eindtoets Basisonderwijs, in het bijzonder de jaargang 2010. In het vervolg zullen we de afkorting EB gebruiken – naast de aanduiding 'Eindtoets' – als we Eindtoets Basisonderwijs bedoelen, gevolgd door een jaartal als we een bepaalde jaargang van de toets op het oog hebben. Deze verantwoording heeft in eerste instantie betrekking op de papieren versie (sinds 2006 is er ook een digitale versie voor alle scholen beschikbaar, waarin ook de eerder digitaal uitgebrachte Inhaaltoets is opgenomen).

Naast deze verantwoording kan men de volgende producten in de beschouwingen betrekken:

Voor de **leerkracht**:

- Handleiding Eindtoets Basisonderwijs
- Handleiding Digitale edities Eindtoets

Op Cito Portal staat:

- Doelenboek (inhoudsverantwoording)
- Brochure 'Interpretatie van het Leerlingrapport'
- Bespreklijst
- Eindtoets 2010 voor Ouders; een selectie van 20 opgaven uit een vorige Eindtoets

Voor de **leerling**:

- Opgavenboekjes Eindtoets 1e, 2e en 3e toetsdag (inclusief Bronnenboekje 2e toetsdag)
- Opgavenboekjes Wereldoriëntatie (indien van toepassing)
- Antwoordbladen 1e, 2e en 3e toetsdag

Voor de **ouders**:

- Ouderkrant Nederlands
- Ouderkrant Turks, Arabisch (indien besteld)

Een Engelstalige versie van de Ouderkrant is te downloaden via de website (www.cito.nl).

De standaard scoringsservice bestaat uit de volgende onderdelen:

- Leerlingrapport (in viervoud)
- Rapportage Eindtoets Basisonderwijs REB (beschikbaar via Cito Portal)

Op het **Leerlingrapport** (zie figuur 4.3) worden de standaardscore en de scores op de hoofdonderdelen Taal, Rekenen-Wiskunde, Studievaardigheden en Wereldoriëntatie gerapporteerd. Per hoofdonderdeel betreft dit het aantal opgaven, het aantal goed en de percentielscore.

Op het Leerlingrapport is verder de zogenoemde poppetjesgrafiek afgebeeld. Deze grafiek geeft grafisch weer wat de betekenis is van de standaardscore voor de keuze van een brugklastype. Bovendien staat in de toelichting boven de poppetjesgrafiek in woorden wat de standaardscore op de Eindtoets zegt over het brugklastype dat het meest in aanmerking komt voor de betreffende leerling.

Met het digitale Rapportageprogramma Eindtoets Basisonderwijs (**REB**) kunnen scholen zelf Schoolrapporten maken op basis van alle mogelijke selecties van leerlingen. Op Schoolrapport A en B staan de gemiddelde standaardscores van de afgelopen drie jaar en de scores per onderdeel. Beide rapporten leveren verder een profiel van de onderscheiden onderdelen. Op basis van het profiel kunnen scholen een discrepantieanalyse uitvoeren. Naast Schoolrapport A en B kunnen scholen met REB ook Leerlingrapporten, Leerlingprofielen (indien scholen het Opgavenboekje Extra hebben afgenomen) en de IC-Tabel opvragen. De afkorting IC staat daarbij voor instroomcorrectie. In de IC-Tabel worden IC-Standaardscores en IC-Percentielen gerapporteerd.

Voor illustraties van deze rapporten kunt u terecht in de Handleiding Eindtoets Basisonderwijs.

Bij deze verantwoording horen als (separate) bijlagen de Handleiding Eindtoets Basisonderwijs 2010, het Doelenboek (inhoudsverantwoording), de brochure Interpretatie van het Leerlingrapport 2010, een exemplaar van de Terugblik en resultaten van de afname van de Eindtoets Basisonderwijs 2010 en exemplaren van de onderscheiden Opgavenboekjes Eindtoets Basisonderwijs 2010.

Het genoemde materiaal maakt een beoordeling van de Eindtoets mogelijk op de volgende aspecten:

- Uitgangspunten van de toetsconstructie
- De kwaliteit van het toetsmateriaal
- De kwaliteit van de handleiding
- Normen
- Betrouwbaarheid
- Validiteit

Deze wetenschappelijke verantwoording heeft met name betrekking op de uitgangspunten van de toetsconstructie (hoofdstuk 2 en 3), de normen (hoofdstuk 4), de betrouwbaarheid en meetnauwkeurigheid (hoofdstuk 5) en de validiteit (hoofdstuk 6). De kwaliteit van het toetsmateriaal en de handleiding is vast te stellen door kennis te nemen van de inhoud van de overige opgesomde producten.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

De Eindtoets is een zogenoemde school- of leervorderingentoets: de toets meet wat een leerling in vergelijking met andere leerlingen in acht jaar basisonderwijs geleerd heeft. De inhoud van de Eindtoets sluit nauw aan bij het onderwijsprogramma voor de basisvaardigheden Taal, Rekenen-Wiskunde en Studievaardigheden. De opgaven zijn een operationalisering van doelstellingen die door het gehele basisonderwijs worden nagestreefd, zogenoemde communale doelstellingen. De toets is geconstrueerd op basis van de verwachting dat deze een bijdrage kan leveren aan een adequate plaatsing van leerlingen in het vervolgonderwijs bij afsluiting van het basisonderwijs. Desgewenst kan aanvullend ook het onderdeel Wereldoriëntatie (eerste afname in 1995) worden afgenomen. Afname van dit onderdeel is facultatief; het wordt niet bij de primaire functie van de toets (namelijk onafhankelijke informatie te geven voor de keuze van een passend brugklatype) betrokken. Sinds 2004 wordt het opgavenboekje Wereldoriëntatie niet meer ieder jaar vernieuwd, maar is het drie jaar op rij te gebruiken (laatste versie 2010-2012).

2.2 Doelgroep

In principe doen alle leerlingen in groep 8 mee aan de afname van de toets. Een uitzondering op deze regel vormen:

- (allochtone) leerlingen die aan het begin van groep 8 vier jaar of korter in Nederland zijn en die het Nederlands onvoldoende beheersen om de opgaven goed te kunnen lezen;
- leerlingen die naar verwachting naar het (voortgezet) speciaal onderwijs of naar het praktijkonderwijs gaan.

Voor de meeste (groepen) leerlingen met beperkingen voor wie de afname van de Eindtoets problemen oplevert, zijn er speciale versies van de toets uitgebracht. Er is een gesproken versie van de Eindtoets Basisonderwijs op cd en daisy-cd uitgebracht, de laatste specifiek geschikt voor de daisyspeler. Vanaf 2003 kunnen scholen ook de vergrote versie (A3-formaat; voor slechtziende leerlingen) en een zwart-witversie (voor kleurenblinde leerlingen) van de opgavenboekjes bestellen. Van de opgavenboekjes van de Eindtoets en Wereldoriëntatie is op aanvraag ook een brailleversie beschikbaar voor blinde leerlingen.

In 2004 is de zogenoemde Niveautoets van start gegaan voor leerlingen met een leerachterstand van tenminste anderhalf jaar op het gebied van Taal én Rekenen. De scores op deze toets kunnen worden omgezet (geëquivalet) naar scores op de reguliere Eindtoets. Sinds 2007 is er een NiveautoetsPlus met een onderdeel technisch lezen en extra opgaven voor spelling en begrijpend lezen.

2.3 Gebruiksdoel

De specifieke functie van de Eindtoets Basisonderwijs is het geven van onafhankelijke informatie voor de keuze van een passend brugklatype. De Eindtoets is een school- of leervorderingentoets: de toets meet wat een kind in vergelijking met andere kinderen in acht jaar basisonderwijs geleerd heeft. En leervorderingen zeggen iets over de vraag hoe goed een leerling de verschillende typen van het voortgezet onderwijs aankan.

Omdat de Eindtoets meestal bij alle leerlingen in groep 8 wordt afgenomen, kan een discrepantieanalyse van de gemiddelde scores op de verschillende onderdelen in het Schoolrapport worden gebruikt bij de evaluatie van het onderwijs. In aanvulling daarop kan de schoolrapportage aangeven wat de plaats van de school is in de landelijke verdeling van schoolgemiddelden en in de verdeling van de schoolgroep waartoe de school behoort.

Met het oog op de optimalisering van onafhankelijkheid van de informatie wordt van de Eindtoets jaarlijks een nieuwe versie geconstrueerd. Alle jaargangen van de Eindtoets zijn equivalent (zie hierover paragraaf 2.4.2.1).

2.4 Theoretische inkadering

2.4.1 Inhoudelijk

Historische achtergrond

Teruggaand in de geschiedenis van de EB, kan worden vastgesteld dat deze zijn oorsprong vindt in de Amsterdamse Schooltoets. Een belangrijk uitgangspunt van de Amsterdamse Schooltoets in 1966 was dat de informatie die de afname van de toets opleverde ook een analyse van de stand of het peil van het onderwijs mogelijk moest maken (Samenwerkende Instituten, RITP en Nutsseminarium voor Pedagogiek, 1967; Deen, 1969). Dat doel had een belangrijke consequentie: de toets zou bij alle leerlingen in het laatste jaar van de basisschool – toen nog Lagere School – moeten worden afgenomen. En niet zoals tot 1966 gebruikelijk was en door wettelijke voorschriften ingegeven, alleen aan leerlingen die in aanmerking kwamen voor onderzoek naar hun mogelijke geschiktheid voor het VHMO, het voorbereidend hoger en middelbaar onderwijs. Aan die voorwaarde moest de inhoud en de moeilijkheid van de toets zijn aangepast. En bovendien: belangrijke informatie voor de analyse van de stand of de kwaliteit van het onderwijs op een school is de vergelijking van de gemiddelde score met die van andere scholen. Daartoe moest de inhoud van de toets aansluiten bij onderdelen van het curriculum waaraan scholen vrijwel zonder uitzondering ongeveer dezelfde aandacht besteden¹.

Cito heeft de ontwikkeling van de Amsterdamse Schooltoets in 1970 overgenomen en heeft aan deze uitgangspunten tot nu toe vastgehouden. Geheel overeenkomstig de uitgangspunten is het onderdeel Wereldoriëntatie facultatief: de inhoud van dit onderdeel verschilt per school (en denominatie) relatief sterk. Sinds 1970 is de opzet van de Eindtoets Basisonderwijs een aantal keren veranderd. In 1973 bijvoorbeeld werd een indirecte meting van de schrijfvaardigheid geïntroduceerd (Wesdorp, 1974), in 1977 werden standaardscores ingevoerd (Jansen, 1976) en werd spelling weer een 'verplicht' onderdeel van de toets. In 2003 zijn veranderingen doorgevoerd die vooral inhoudelijk van aard zijn. Begrijpend lezen werd ondergebracht bij het onderdeel Taal in plaats van bij Informatieverwerking. De naamgeving van het onderdeel Informatieverwerking werd vanaf 2003 veranderd in Studievaardigheden. Aan dit onderdeel werd een nieuwe rubriek toegevoegd: 'Hanteren van studieteksten'². Het onderdeel Taal werd uitgebreid met het onderdeel 'Woordenschat' (20 opgaven).

Begripsmatige achtergrond

Zou zo'n toets ook (of juist bij uitstek) geschikt zijn voor het primaire doel? Hoofdstuk 6, over validiteit, zal laten zien dat de Eindtoets goed aan zijn doel beantwoordt. Is daar ook een inhoudelijke verklaring voor? Is het logisch dat een schoolvorderingentoets waarin de resultaten van eerder onderwijs worden geëvalueerd iets zegt over latere leerresultaten? Het lijkt gerechtvaardigd om te veronderstellen dat leerprestaties uit het verleden iets zeggen over leerprestaties in de toekomst. Verschillen in leerprestaties op de onderdelen die in de Eindtoets vertegenwoordigd zijn, kunnen vermoedelijk voor een belangrijk deel worden verklaard door verschillen in eigenschappen als 'intelligentie', 'concentratie', 'motivatie' en 'doorzettingsvermogen'. Via de Eindtoets worden deze eigenschappen indirect en afgewogen³ gemeten. Zo bezien zou de Eindtoets wel eens veelzijdiger kunnen zijn dan op grond van oppervlakkige waarneming vaak gedacht wordt.

Uit een recent onderzoek van Van Dijk en Tellegen (2004) komen aanwijzingen voor de houdbaarheid van de veronderstelling hierboven, althans voor zover het om 'intelligentie' gaat. Zij vinden tussen de intelligentietest die zij verantwoorden (NIO) en de Eindtoets hoge tot zeer hoge correlaties, zowel met de onderdelen van de Eindtoets als met de totaalscore (met deze laatste een correlatie van .78; er werd een correlatie gevonden van .75 met het advies dat is gebaseerd op de Eindtoets). Ook de totaalscore van een

¹ Een andere consequentie van de uitgangspunten van de Amsterdamse Schooltoets is dat er strikt genomen meer opgaven in de EB zijn opgenomen dan voor de primaire functie noodzakelijk is. Door een goede spreiding van de afname zal deze consequentie eerder positief dan negatief bijdragen aan de validiteit en betrouwbaarheid.

² Het aantal opgaven Studieteksten bedraagt 10. Met dit aantal werd het aantal opgaven Begrijpend lezen verlaagd (van 40 naar 30 opgaven).

³ In de Ouderkrant bij de Eindtoets Basisonderwijs wordt gesproken van een 'persoonlijke mix'.

van de Eindtoets afgeleide toets in klas 2 en 3 van het voortgezet onderwijs, zo blijkt uit het onderzoek van Van Dijk en Tellegen, correleert in hoge mate met de intelligentiescore op de NIO (respectievelijk .78 en .82). Eigen onderzoek ondersteunt deze bevinding: in 2003 beschikten we over de score van 1351 leerlingen op de ISI-Intelligentietest en over hun score op de Entreetoets groep 7. We vonden een correlatie van .72. Gezien de gelijkenis in concept tussen Entreetoets en Eindtoets en gezien de zeer hoge correlaties die we steeds vinden tussen Entreetoets groep 7 en Eindtoets is het plausibel dat de correlatie tussen ISI en Eindtoets van dezelfde orde van grootte zal zijn. In 2009 werd deze aanname empirisch onderbouwd (zie Van Boxtel en Hemker, 2009). Sedert dat jaar is het mogelijk om de Cito Intelligentietest, die een bewerking vormt van de voormalige ISI-Intelligentietest, af te nemen naast de Eindtoets. De correlatie tussen de betreffende intelligentie-index en de totaalscore op de Eindtoets bedroeg .75 (.72 voor de totaalscore exclusief Wereldoriëntatie). Deze bevindingen tonen aan dat Eindtoets en intelligentiemetingen ergens tussen de 50% en de 60% van hun variantie gemeenschappelijk hebben. Dit is een comfortabel resultaat in twee opzichten: de correlatie is enerzijds substantieel, en toont aan dat schoolse prestaties inderdaad samenhangen met intelligentie (waar een nul-correlatie toch wel bevreemding zou wekken), maar anderzijds is de correlatie laag genoeg om een zelfstandig gebruik van de Eindtoets te rechtvaardigen. In hoofdstuk 6 gaan we nader in op de samenhangen tussen intelligentietests en leervorderingentoetsen enerzijds en doorstroomadviezen en feitelijke doorstroomgegevens anderzijds. Vooruitlopend hierop kan worden gesteld dat de voorspellende waarde van leervorderingentoetsen in vergelijking met intelligentietests wat hoger ligt. Dat geldt zowel voor correlaties met doorstroomadviezen van de leerkracht als voor de feitelijke doorstroming, uiteraard op voorwaarde dat de leervorderingentoets een deugdelijke afspiegeling vormt van de onderwijsdoelstellingen. Vergelijk in dit opzicht bijvoorbeeld de studie van Boland en Mommers (1987) waarin het ISI-prestatiedeel (intelligentie- en leervorderingen) en de Eindtoets met elkaar werden vergeleken. Hun regressieanalyses lieten zien dat leervorderingentoetsen stevast als beste voorspeller naar voren komen en dat – in termen van verklaarde variantie in schoolsucces⁴ – intelligentietests geen toegevoegde waarde hebben. Dat leervorderingentoetsen in dit opzicht superieur zijn, is vermoedelijk te danken aan het feit dat leervorderingentoetsen niet alleen iets zeggen over de intelligentie van de leerling (zie de boven besproken correlaties), maar indirect ook over persoonlijke eigenschappen die het leereffect mede bepalen. Het is waarschijnlijk dat deze eigenschappen in scores op intelligentietests maar zeer ten dele tot uitdrukking komen. Deze persoonlijke eigenschappen bepalen niet alleen mee hoeveel een leerling, gegeven zijn intelligentie, in het basisonderwijs opsteekt, maar zij zijn ook van invloed op wat het in het voortgezet onderwijs aankan. Hun invloed doet zich dus zowel gelden in de predictor als in het criterium.

Onderdelen

Binnen de Eindtoets kunnen we op grond van de inhoud een drietal subtoetsen onderscheiden: Taal, Rekenen-Wiskunde en Studievaardigheden. Daarnaast is er het (facultatieve) onderdeel Wereldoriëntatie. Deze hoofdgebieden zijn op hun beurt opgedeeld in rubrieken. In hoofdstuk 3 zullen we deze onderdelen in meer detail beschrijven en verantwoorden. Voor dit moment past de vraag hoe met de opzet van de constructie van de Eindtoets zo goed mogelijk tegemoet gekomen wordt aan de uitgangspunten en functies ervan. In dit opzicht zijn vooral de maatregelen van belang die de afstemming op communale inhoud van het onderwijs borgen; immers, zoals hierboven al is aangegeven, voorwaarde voor een optimaal voorspellend vermogen is dat de toetsinhoud een adequate afspiegeling vormt van deze communale onderwijsinhouden:

- Cito analyseert doelstellingen en methoden en gaat aan de hand van de resultaten na of de Eindtoets nog goed aansluit bij de inhoud van het onderwijs.
- Cito publiceert een inhoudsverantwoording, het Doelenboek (Staphorsius, 2003, 2009) en stelt deze zo nodig bij. In het Doelenboek staan doelstellingen en voorbeeldopgaven. Alle deelnemende scholen ontvangen dit Doelenboek.
- Een onderdeel van de jaarlijkse proeftoets is het verzoek om commentaar van leerkrachten.
- De constructie van opgaven vindt plaats in commissies waarvan de leden ervaring hebben in het basisonderwijs.

⁴ Boland en Mommers gebruiken als operationalisatie van schoolsucces de plaatsing in het voortgezet onderwijs.

Verder is het belangrijk om vast te stellen dat de Eindtoets een glazen huis is. Ongeveer 6200 scholen nemen de Eindtoets af. Scholen bespreken de gemiddelde scores met het bevoegd gezag. We kunnen er dan ook van overtuigd zijn dat scholen de keuze van niet passende inhoud niet onbecommentarieerd zullen laten. De laatste jaren heeft Cito geen commentaar bereikt over de relatie tussen de doelstellingen van het onderwijs met betrekking tot de in de Eindtoets vertegenwoordigde onderdelen enerzijds en de opgaven in de toets anderzijds.

In hoofdstuk 3 gaan we uitgebreider in op de inhoud en structuur van de Eindtoets. Daarbij zal het meest recente Doelenboek van 2009 het uitgangspunt zijn.

2.4.2 Psychometrisch

In deze paragraaf wordt allereerst beschreven hoe Cito bewerkstelligt dat elke nieuwe versie van de Eindtoets een equivalente versie van de voorafgaande is. Dat vereist zorgvuldige constructieprocedures die in paragraaf 2.4.2.1 uitvoerig aan de orde komen. In deze paragraaf zal ook duidelijk worden dat het gehanteerde IRT-meetmodel in deze procedures een cruciale rol speelt. In paragraaf 2.4.2.2 wordt uitvoerig op dit meetmodel ingegaan.

Deze paragrafen zijn juist daarom zo belangrijk omdat methodologisch en empirisch onderbouwd wordt dat de Eindtoets, ondanks de jaarlijkse ingrijpende vernieuwing van de inhoud, steeds in essentie dezelfde toets blijft. Het bewijs voor deze stellingname is te ontleen aan de resultaten van onderzoek naar de jaargangen 2003 tot en met 2006 dat eerder is gepubliceerd in Staphorsius, Béguin, Engelen, Kleintjes & Verhelst (2007). In hoofdstuk 6 komen we hierop terug en presenteren we onder de noemer validiteit soortgelijke en andere onderzoeksresultaten met betrekking tot de Eindtoets 2010. Daar zal onder meer duidelijk worden dat de conclusies met betrekking tot equivalentie ook en nog steeds gelden voor jaargang 2010.

De implicatie hiervan is, dat studies die in het verleden zijn uitgevoerd naar de eigenschappen en de validiteit van de Eindtoets hun geldigheid behouden, ook voor de hier verantwoorde versie van 2010. In hoofdstuk 6 zullen resultaten van enkele eerdere studies in samengevatte vorm worden gepresenteerd, voor zover zij een bijdrage leveren aan de empirische onderbouwing van de validiteit.

2.4.2.1 Constructie- en equivaleringsprocedures

Zoals gezegd bestaat de Eindtoets ieder jaar volledig uit nieuw itemmateriaal. De primaire functie van de toets (het geven van een onafhankelijk advies) vereist maatregelen die garanderen dat het construct dat wordt gemeten door elk van de toetsonderdelen constant blijft. Deze constantie heeft niet zozeer betrekking op de moeilijkheid van de items als wel op hun homogeniteit met betrekking tot het construct dat ze pretenderen te meten.

Proeftoetsonderzoek

Een belangrijk instrument dat Cito met het oog op de continuïteit hanteert, is het zogenaamde proeftoetsonderzoek. Daarin wordt in januari van elk jaar elke conceptopgave aan ongeveer 300 tot 400 leerlingen voorgelegd. Het proeftoetsonderzoek van 2009 bijvoorbeeld, bevatte itemmateriaal dat

- deel uitmaakt van de EB 2010 (de hier verantwoorde jaargang);
- deel zal uitmaken van de EB 2011;
- diende om nieuwe itemformaten uit te testen, die eventueel in latere EB's kunnen worden gebruikt.

Het aantal items dat wordt geproeftoetst, is een veelvoud van het aantal dat uiteindelijk in de EB wordt ingezet. Er moet immers rekening worden gehouden met uitval, bijvoorbeeld wegens meer of minder triviale fouten in de constructie of extreme moeilijkheid of gemakkelijkerheid. Tezelfdertijd ontstaat er op deze manier een overschot aan kwalitatief goede opgaven, zodat voor elk onderdeel gaandeweg itembanken ontstaan, waaruit kan worden geput voor een volgende jaargang van de Eindtoets. Een belangrijk kenmerk van deze opgavenbanken is dat ze gekalibreerd zijn met een unidimensionaal IRT-model (OPLM; Verhelst en Eggen, 1989; zie verder paragraaf 2.4.2.2), waardoor niet alleen de psychometrische kenmerken (parameters) van de opgaven worden geschat, maar waarbij tevens wordt nagegaan of de opgaven van een onderdeel kunnen worden beschreven met een unidimensionale onderliggende vaardigheid. Een gedetailleerde

verantwoording van deze werkwijze voor bijvoorbeeld het onderdeel Taal is te vinden in Staphorsius, Verhelst en Kleintjes (2001) en in Staphorsius, Krom, Kleintjes en Verhelst (2001). Staphorsius en Verhelst (2001) geven tevens een uitvoerige beschrijving van de gevolgde werkwijze voor de onderdelen van de Toelatingstoets, die op dezelfde manier zijn gekalibreerd en onderzocht.

Opgavenbanken

Voor het samenstellen van toetsen voor het primair onderwijs beschikt Cito over opgavenbanken. Die liggen ten grondslag aan onder meer de toetsen in het Cito Leerling- en onderwijsvolgsysteem (LOVS), de Entreetoetsen en de Eindtoets Basisonderwijs. Een opgavenbank is nadrukkelijk niet eenvoudigweg een verzameling opgaven of items waaruit een toetsconstructeur min of meer naar willekeur een aantal items selecteert om een nieuwe toets te construeren. Hieronder wordt beschreven wat de vereisten zijn om van een deugdelijke en psychometrisch goed gefundeerde opgavenbank te kunnen spreken.

Unidimensionaal continuüm en latente vaardigheid

Het algemene uitgangspunt is dat de vaardigheden die in de onderdelen van de Eindtoets worden gemeten kunnen worden opgevat als een unidimensionaal continuüm (de reële lijn), en dat elke leerling voorgesteld kan worden als een punt op die lijn, met andere woorden: als een getal. Het getal drukt de mate van vaardigheid uit, waarbij een groter getal wijst op een grotere vaardigheid. Het doel van de meetprocedure – het afnemen van de toets – is de plaats van de leerling op dit continuüm zo nauwkeurig mogelijk te bepalen. De uitkomst van de meetprocedure bestaat strikt genomen uit twee grootheden: de eerste is de schatting van de plaats van de leerling op het vaardigheidscontinuüm. De tweede grootheid geeft aan hoe nauwkeurig die schatting is, en heeft dus de status van een standaardfout, te vergelijken met de standaardmeetfout uit de klassieke testtheorie. De antwoorden van een leerling op de items worden beschouwd als indicatoren van de vaardigheid, hetgeen ruwweg betekent dat men verwacht dat alle items in de bank deze zelfde vaardigheid meten. De vaardigheid zelf wordt als niet-observeerbaar beschouwd, en daarom gewoonlijk omschreven als een latente vaardigheid.

'Moeilijkheid' in de Item Respons Theorie

Hoewel items dezelfde vaardigheid meten, kunnen ze toch systematisch van elkaar verschillen. Het belangrijkste verschil tussen de items is hun moeilijkheidsgraad. In de klassieke testtheorie wordt de moeilijkheidsgraad uitgedrukt met een zogenaamde p-waarde, de proportie correcte antwoorden op het item in een welbepaalde populatie van leerlingen. In de Item Respons Theorie (IRT) die voor het construeren van de opgavenbanken wordt gebruikt, hanteert men echter een andere definitie van moeilijkheid: ruwweg gesproken is het de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden. Dit verschil in definitie van de moeilijkheidsgraad tussen klassieke theorie en IRT is uitermate belangrijk: men kan verwachten dat de p-waarde van een item in groep 8 groter zal zijn dan in groep 6, waardoor duidelijk wordt dat de p-waarde een relatief begrip is: ze geeft de moeilijkheid aan van een item in een bepaalde populatie. Binnen de IRT is de moeilijkheid van een item gedefinieerd in termen van de onderliggende vaardigheid, zonder enige verwijzing naar een bepaalde populatie van leerlingen. Zo kan men ook de uitspraak begrijpen dat in de IRT vaardigheid en moeilijkheid op eenzelfde schaal liggen.

Kansmodel

De ruwe omschrijving van de moeilijkheidsgraad die in de vorige alinea werd gehanteerd (de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden) behoeft enige verdere uitwerking. Men zou deze omschrijving kunnen opvatten als een drempel: heeft een leerling die mate van vaardigheid niet, dan is hij niet in staat het item juist beantwoorden; heeft hij die drempel wel gehaald, dan geeft hij (gegarandeerd) het juiste antwoord. Deze interpretatie weerspiegelt een deterministische kijk op het antwoordgedrag van de leerling, die echter in de praktijk geen stand houdt, omdat eruit volgt dat een leerling die een moeilijk item correct beantwoordt geen fout kan maken op een gemakkelijk item. Daarom wordt in de IRT een kansmodel gebruikt: hoe groter de vaardigheid, des te groter de kans dat een item juist wordt beantwoord. De moeilijkheidsgraad van een item wordt dan gedefinieerd als de mate van vaardigheid die nodig is om met een kans van precies een half, een juist antwoord te kunnen produceren (zie verder ook de volgende paragraaf over het meetmodel).

Kalibratie

In het voorgaande zijn nogal wat veronderstellingen ingevoerd (unidimensionaliteit; alle items zijn indicatoren voor dezelfde vaardigheid; kansmodel) die niet zonder meer voor waar kunnen worden aangenomen; er moet aangetoond worden dat al die veronderstellingen deugdelijk zijn. Dit 'aantonen' gebeurt met statistische gereedschappen waarop later nog dieper in wordt gegaan. Maar vóór de items in een toets gebruikt kunnen worden, moet ook geprobeerd worden de waarden van de moeilijkheidsgraden te achterhalen. Dit gebeurt met een statistische schattingsmethode die wordt toegepast op de itemantwoorden die bij een steekproef van leerlingen zijn verzameld. Het hele proces van moeilijkheidsgraden schatten en verifiëren of de modelveronderstellingen houdbaar zijn, wordt kalibratie of ijking genoemd. De steekproef van leerlingen (in de boven al aangeduide proeftoets) die hiervoor wordt gebruikt heet kalibratiesteekproef.

Afnamedesigns

Meestal bevat een opgavenbank meer items dan een doorsnee toets, zodat het praktisch niet doenbaar is om alle items aan alle leerlingen voor te leggen. Elke leerling in de kalibratiesteekproef krijgt derhalve slechts een (klein) gedeelte van de items uit de opgavenbank voorgelegd. Er is dan sprake van een zogenoemd onvolledig design. Dit gedeeltelijk voorleggen moet met de nodige omzichtigheid gebeuren. Verderop wordt ingegaan op het afnamedesign dat voor de kalibratie is gebruikt, de geïnteresseerde lezer wordt verwezen naar Eggen (1993).

Implicaties van gekalibreerde opgavenverzameling

Als de kalibratie met succes uitgevoerd is, is het resultaat een zogenoemde gekalibreerde itembank. In het kalibratieproces worden de items die niet passen bij de verzameling uit de collectie verwijderd.

De opgavenbank bevat voor elk item niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen, en de statistische zekerheid dat alle items dezelfde vaardigheid aanspreken.

Meer over de kalibreringsprocedure en een bespreking van de resultaten daarvan voor de Eindtoets 2010 is te vinden in hoofdstuk 4 over de normering van de Eindtoets.

Equivalering en de structuur van de Eindtoets

Hoe verhouden de verschillende onderdelen van de EB zich tot elkaar en hoe zorgt Cito ervoor dat deze verhoudingen van jaar tot jaar equivalent blijven?

Door gebruik te maken van een IRT-model kunnen de correlaties tussen de constructen die de verschillende onderdelen meten, worden geschat. In dit model wordt een multivariaat-normale verdeling van de constructen verondersteld. De parameters van deze multivariaat-normale verdeling worden geschat met speciaal daartoe bij Cito ontwikkelde software.

Tabel 2.1 Latente correlaties tussen onderdelen van de EB (2004, 2005)

	Rekenen-Wiskunde	Begrijpend lezen	Schrijven	Spelling	Woordenschat	Studievaardigheden
Rekenen-Wiskunde	1	.720	.747	.634	.592	.856
Begrijpend lezen	.719	1	.932	.669	.833	.894
Schrijven	.710	.927	1	.740	.755	.904
Spelling	.600	.667	.729	1	.551	.682
Woordenschat	.600	.785	.701	.508	1	.735
Studievaardigheden	.867	.889	.876	.652	.725	1

In tabel 2.1 (ontleend aan Staphorsius, Béguin, Engelen, Kleintjes & Verhelst (2007)) zijn bij wijze van voorbeeld de geschatte correlaties weergegeven voor de Eindtoets 2004 (boven de diagonaal) en de Eindtoets 2005 (onder de diagonaal). Dat hadden even goed twee willekeurige andere jaargangen van de Eindtoets kunnen zijn. De correlaties onder en boven de diagonaal lijken steeds heel erg op elkaar. De twee vetgedrukte getallen staan in de cellen waar de correlaties voor de twee gerapporteerde jaren het sterkst van elkaar verschillen, maar ook daar is het verschil verwaarloosbaar, namelijk .054⁵. De drie onderdelen van Rekenen-Wiskunde (Getallen en bewerkingen; Verhoudingen, breuken en procenten; Meten, meetkunde, tijd en geld) correleren onderling hoger dan .97 en worden bij de equivalering als een homogene schaal behandeld. Voor de onderdelen van Studievaardigheden geldt dat de intercorrelaties hoog zijn, maar minder hoog dan bij Rekenen-Wiskunde (tussen .80 en .90). De exacte waarden zijn in dit kader wat minder interessant; voor deze getallen verwijzen we naar Staphorsius et al. (2007). Voor het facultatieve onderdeel Wereldoriëntatie kan opgemerkt worden dat de onderscheiden deelgebieden onderling minder samenhang vertonen. De intercorrelaties liggen tussen .75 en .85.

Uit de tabel blijkt dat er van jaar tot jaar een behoorlijk stabiele structuur van intercorrelaties tussen de onderdelen bestaat, ook als we het onderdeel Taal opsplitsen in deelvaardigheden. Dat geldt niet alleen voor de twee jaargangen die hierboven als voorbeeld zijn gekozen, maar voor alle versies van de Eindtoets. In hoofdstuk 6 kan men een soortgelijke tabel vinden voor de jaargangen 2009 en 2010. Ook de intercorrelaties die in aanvulling op tabel 2.1 werden besproken laten over de jaren heen een stabiel beeld zien.

Daarnaast is het duidelijk dat de totaalscore op de Eindtoets niet zonder meer kan worden geïnterpreteerd als een homogene testscore, maar duidelijk als een mengsel van verschillende relevante vaardigheden. Bij het hanteren van zo'n mix kan men zich twee belangrijke vragen stellen:

- Wat is de rechtvaardiging van de onderlinge verhoudingen van het aantal opgaven in de drie hoofdonderdelen (Taal, 100; Rekenen-Wiskunde, 60; Studievaardigheden, 40)? Deze vraag is zeker niet onbelangrijk met het oog op de inhoudsvaliditeit. Er is gekozen voor een evenwichtige verdeling van opgaven die recht doet aan de aandacht voor deze onderdelen in het onderwijs. De verdeling is en wordt steeds afgestemd op het onderwijsveld: een grote onbalans (bijvoorbeeld met bijna uitsluitend aandacht voor Rekenen-Wiskunde) zou zeker een onevenwichtig beeld geven van wat het onderwijs te bieden heeft. Men zou in dat geval vanuit het veld protest en tegenstand mogen verwachten.

⁵ Let wel dat de gerapporteerde waarden schattingen zijn van de correlaties tussen de latente variabelen. Ze zijn dus niet gedrukt (geattenuëerd) door de onbetrouwbaarheid van de metingen.

Dit heeft zich echter nooit voorgedaan, ook niet na de verzwaring van de talige component met de rubriek Woordenschat in 2003. Het is van belang om te vermelden dat de keuze van deze mix boven andere mogelijke mixen gevolgen heeft voor bepaalde vergelijkingen die voor het onderwijsbeleid van belang kunnen zijn. Zo blijkt, in 2010 opnieuw, dat de prestaties van de meisjes gemiddeld iets lager zijn dan die van de jongens (zie ook hoofdstuk 6), hoewel er bij de jongens relatief meer leerlingen zijn met een leerachterstand (zittenblijvers in of vóór groep 8). Uitsplitsing naar de drie hoofdonderdelen van de Eindtoets toont aan dat meisjes beter presteren dan jongens op het onderdeel Taal. Op het onderdeel Studievaardigheden is sprake van een verwaarloosbaar verschil tussen jongens en meisjes, maar op Rekenen-Wiskunde presteren meisjes duidelijk zwakker dan jongens. Door de gehanteerde relatieve gewichten in de totaaltoets valt de totale score licht in het nadeel van meisjes uit, maar het zal duidelijk zijn dat een andere weging (met bijvoorbeeld minder gewicht voor Rekenen-Wiskunde) de balans de andere richting kan doen uitslaan. Ten slotte is er een overweging om – zelfs als men het met de gehanteerde weging oneens is – die niet lichtvaardig te veranderen. Die overweging heeft te maken met de tweede vraag:

- Hoe beïnvloedt de gekozen weging de constructvaliditeit van de Eindtoets? Juist het constant houden van de relatieve weging van de onderdelen samen met de voorzorgsmaatregelen die men treft voor het behoud van hetzelfde construct binnen de onderdelen zorgt ervoor dat het construct van de Eindtoets als geheel van jaar tot jaar stabiel blijft.

Equivalering en de vergelijkbaarheid van scores

Er is niet alleen de noodzaak van het constant houden van het construct over de tijd, maar er blijft nog een vraagstuk over. We willen immers niet alleen dat verschillende jaargangen van de Eindtoets hetzelfde construct meten, maar ook dat de behaalde scores door de jaren heen vergelijkbaar zijn: een standardscore van 538 hoort als kwantificering van het construct dat door de Eindtoets wordt gerepresenteerd hetzelfde te betekenen in 2005 als in 2010.

We bespreken hier schematisch op welke manier de equivalering tussen 2003 en 2006 verliep. We kiezen hier, wederom volstrekt willekeurig, de equivalering van de Eindtoets 2005 en 2006 als voorbeeld (gegevens opnieuw ontleend aan Staphorsius, Béguin, Engelen, Kleintjes & Verhelst (2007)). Resultaten voor de jaargang 2010 bewaren we voor hoofdstuk 6. Daarbij zullen we ook laten zien dat de equivaleringsprocedures inmiddels nog wat verder zijn aangescherpt.

Omdat de jaargangen van de EB geen opgaven gemeenschappelijk hebben is het onmogelijk om zonder toegevoegde informatie of veronderstellingen deze toetsen met elkaar te vergelijken. De toegevoegde informatie halen we uit de proeftoetsen (PT) die een jaar eerder in 2004, respectievelijk 2005 zijn gehouden. Schematisch ontstaat dan de situatie zoals afgebeeld in figuur 2.1. Elke rij stelt een groep leerlingen voor en elke kolom een aantal opgaven. Een dergelijk design geldt voor elk onderdeel van de Eindtoets dat afzonderlijk wordt gekalibreerd. Het aantal leerlingen in de groepen EB 2005 en EB 2006 stelt geen enkel probleem: destijds namen tussen de 150 000 en 165 000 leerlingen deel aan de EB en al deze leerlingen maakten alle opgaven.

Voor de proeftoetsen liggen de zaken echter veel moeilijker: er worden veel meer dan 200 opgaven ingezet en het is niet verantwoord (en ook niet haalbaar) om alle leerlingen alle items te laten maken. We moeten bovendien tevreden zijn met steekproeven van bescheidener omvang dan de grote aantallen leerlingen die bij de afname van de EB gebruikelijk zijn. Er worden doorgaans ongeveer 600 antwoorden per opgave verzameld. Het belangrijkste is echter dat door het gebruik van het design in figuur 2.1 de twee jaargangen van de EB met elkaar vergeleken kunnen worden: de groepen PT 2004 (a) en PT 2005 (a) hebben allebei opgaven gemaakt die tot de twee jaargangen behoren, en dit levert precies de mogelijkheid op om de twee toetsen (of het relevante onderdeel) op eenzelfde schaal te kalibreren. Met andere woorden: de parameters van alle items van beide Eindtoetsen kunnen nu op basis van deze informatie worden geschat.

Figuur 2.1 Beschikbare informatie voor het equivaleren van twee jaargangen EB (2005 en 2006 bij wijze van voorbeeld; vereenvoudigd)

EB 2005						
EB 2006						
PT 2004 (a)						
PT 2004 (b)						
PT 2004 (c)						
PT 2005 (a)						
PT 2005 (b)						
PT 2005 (c)						

Deze informatie, samen met de informatie over de latente correlaties tussen de onderdelen van de toets, maakt het mogelijk om een fictieve steekproef samen te stellen uit de populatie die deelgenomen heeft aan EB 2005. Deze fictieve steekproef ‘maakt’ zowel de toetsen EB 2005 als EB 2006. Dit levert een schatting op van de scoreverdeling van de populatie EB 2005 op de toetsen EB 2005 en EB 2006⁶. Deze twee verdelingen, en meer in het bijzonder hun gemiddelde en hun standaardafwijking, worden gebruikt voor een lineaire equivalering van beide jaargangen. Deze laatste stap is als volgt schematisch samen te vatten: met X_{2005} stellen we de score op de EB 2005 voor van een fictieve maar random getrokken persoon uit de verdeling (populatie) van de EB 2005. Deze score is een getal in het bereik (0-200). In 2005 is een lineaire transformatie gehanteerd die geobserveerde scores omzet in standaardscores (Stsc; deze liggen in het bereik 501-550)⁷. Voor leden van de populatie 2005 hanteren we dus de transformatie

$$Stsc_{2005} = A_{2005} \times X_{2005} + B_{2005}$$

Voor de fictieve steekproef van 150 000 leerlingen berekenen we de gemiddelde standaardscore $Stsc_{2005}$ en de standaardafwijking $SD(Stsc_{2005})$. We krijgen ook (fictieve) geobserveerde scores (X_{2006}) als we aan de steekproef de EB 2006 voorleggen. Deze scores moeten ook worden omgezet in standaardscores met de formule

$$Stsc_{2006} = A_{2006} \times X_{2006} + B_{2006}$$

waarbij A_{2006} en B_{2006} echter niet gekend zijn. Lineaire equivalering betekent dan dat A_{2006} en B_{2006} zo worden gekozen dat

$$Stsc_{2005} = Stsc_{2006} \text{ en } SD(Stsc_{2005}) = SD(Stsc_{2006})$$

⁶ De omvang van de fictieve steekproef voor dat jaar is 150 000.

⁷ In feite is het proces nog ingewikkelder: eerst wordt de geobserveerde score via een lineaire transformatie omgevormd tot een getransformeerde score die voor de meeste (maar niet alle) leerlingen in het interval (501-550) ligt. Vervolgens wordt deze getransformeerde score afgerond tot een geheel getal en uitkomsten die kleiner zijn dan 501 of groter dan 550 worden tot respectievelijk 501 en 550 teruggebracht. Dit levert de scores op die worden meegedeeld aan scholen en ouders. Alle berekeningen zijn echter uitgevoerd op de niet afgeronde getransformeerde scores.

Let wel dat met deze methode op geen enkele manier gebruik wordt gemaakt van de empirische verdeling van de scores op EB 2006. Het is zelfs zo dat in principe de equivalering van EB 2005 en EB 2006 kan worden uitgevoerd voordat de EB 2006 is afgenomen (zij het met een design dat ingewikkelder is dan het design dat in figuur 2.1 is afgebeeld). De empirische distributie van de scores in 2006 leert dan of het gemiddelde toe- of afgenomen is ten opzichte van het voorgaande jaar.

Uiteraard kleeft er aan de uitkomsten van de analyse hierboven een statistische fout, de equivaleringsfout. De grootte daarvan is afhankelijk van de grootte van de steekproef en van het aantal ankeropgaven, dat wil zeggen het aantal opgaven uit bijvoorbeeld de verschillende jaargangen die door eenzelfde leerling worden gemaakt. Het schatten van de standaardfout van de equivalering is gezien de complexiteit van de gebruikte modellen een bijzonder moeilijke aangelegenheid. Om toch enige grip te krijgen op de nauwkeurigheid van de equivalering wordt de volgende procedure toegepast.

We gaan uit van de veronderstelling dat er in de loop van een gering aantal jaren geen dramatische veranderingen plaatsvinden in het landelijke algemene niveau van de leervorderingen. Meer specifiek gaan we ervan uit dat er een substantieel aantal aan de EB deelnemende scholen is waarvan het niveau vrijwel constant is. Om zulke scholen te detecteren worden de volgende criteria gebruikt:

- De scholen nemen in het lopende jaar deel aan de Eindtoets en deden dit ook in de drie voorgaande jaren. Voor de periode 2003-2006 bijvoorbeeld betrof dit 5805 scholen.
- De scholen zijn redelijk constant wat betreft het aantal deelnemende leerlingen. Als maat wordt genomen de variantie van het aantal deelnemende leerlingen in het lopende jaar en de drie voorgaande jaren.
- De school is redelijk constant wat betreft gemiddelde prestatie op de Eindtoets. Als maat wordt genomen de variantie van de gemiddelde standardscore over de drie voorgaande jaren.
- Als maat om de scholen te rangordenen wordt de som van de twee hierboven besproken varianties genomen.

Vervolgens worden vier groepen scholen gevormd: de eerste groep is de groep van 600 scholen met de allerkleinste varianties, verder aangeduid als groep 0; de drie andere groepen van 600 scholen worden random samengesteld uit de 1800 scholen met de kleinste varianties (deze groepen worden aangeduid als groep 1, 2 en 3). Groep 0 is daarbij gelijkelijk verdeeld over de groepen 1, 2 en 3.

Aannemende dat het gemiddeld niveau in een groep constant is gebleven kunnen de scores van twee eindtoetsen lineair met elkaar worden geëquivalerd. In tabel 2.2 staat een overzicht van de gemiddelde standardscores en hun standaardafwijkingen die we zouden hebben verkregen met deze methode bij het equivaleren van Eindtoets 2004, 2005 en 2006 met de Eindtoets van 2003 als referentie. In de laatste twee rijen staat het resultaat als de groepen 1, 2 en 3 worden gecombineerd.

De overeenkomsten zoals deze tot uitdrukking komen in de tabel zijn in het algemeen zeer goed te noemen. De kwaliteit van de toegepaste procedures lijkt dus hoog. Dit is ook belangrijk omdat de adviezen voor de keuze van een brugklatype die Cito op basis van de scores op de Eindtoets verstrekt, zijn gebaseerd op een retrospectief onderzoek naar de toelating en doorstroom van leerlingen in het voortgezet onderwijs. Het onderzoek maakt het mogelijk een relatie te leggen tussen de standardscore van een leerling en het school- of brugklatype waarop de leerling de grootste kans op succes heeft. Het behoeft geen betoog dat de kwaliteit van de equivalering bijdraagt aan de predictieve validiteit van de EB: het Toelatings- en doorstroomonderzoek in bijvoorbeeld 2005 gaat de toelating en doorstroom na van leerlingen die in 2003 en 2004 de EB maakten en moet de standardscores van 2006 van een adequate betekenis voorzien.

Tabel 2.2 *Vergelijking van uitkomsten van de equivaleringsprocedures*

	parameter	2004	2005	2006
'officieel'	gemiddelde	535.16	534.55	535.00
	SD	9.70	10.07	9.92
sample 0	gemiddelde	535.08	534.82	534.96
	SD	9.72	9.75	9.64
sample 1	gemiddelde	535.01	534.77	535.16
	SD	9.56	9.80	9.69
sample 2	gemiddelde	535.26	534.95	535.47
	SD	9.54	9.94	9.72
sample 3	gemiddelde	535.03	534.93	535.04
	SD	9.49	9.91	9.71
sample 1+2+3	gemiddelde	535.10	534.88	535.22
	SD	9.53	9.88	9.71

De uitkomsten van de analyses die in paragraaf 2.4.2.1 zijn gerapporteerd, maken duidelijk dat de verschillende versies van de Eindtoets van jaar tot jaar hun structuur behouden in termen van intercorrelaties tussen de onderdelen. Daarmee wordt het construct dat in de Eindtoets wordt geoperationaliseerd constant gehouden. De equivaleringsprocedures zorgen er daarnaast voor dat de gemiddelde score en variantie daarin van jaar tot jaar vrijwel exact hetzelfde blijven. Soortgelijke gegevens over de equivalentie van de Eindtoets 2010 ten opzichte van eerdere versies worden besproken in hoofdstuk 6 over validiteit. Daarin komen ook de resultaten van het genoemde Toelatings- en doorstroomonderzoek aan de orde.

2.4.2.2 Het gehanteerde meetmodel

In de Eindtoets Basisonderwijs is gebruikgemaakt van een op de itemresponstheorie (IRT) gebaseerd meetmodel zoals dat bij Cito gebruikelijk is. Dergelijke modellen verschillen in een aantal opzichten nogal sterk van de klassieke testtheorie (Verhelst, 1993; Verhelst & Kleintjes, 1993; Verhelst & Glas, 1995). Bij de klassieke testtheorie staan de toets en de toetsscore centraal. Het theoretisch belangrijkste begrip in deze theorie is de zogenoemde ware score, de gemiddelde score die de persoon zou behalen indien de test een oneindig aantal keren onder dezelfde condities zou worden afgenomen. Die notie geeft een van de belangrijkste (praktische) obstakels van deze theorie voor ons onderzoek weer: het is problematisch om toetsscores te vergelijken die verkregen zijn in een onvolledig design. Hoewel er methoden bestaan binnen de klassieke testtheorie om toetsscores te equivaleren (Engelen & Eggen, 1993), schiet deze benadering tekort als het gaat om de centrale vraag: hoe wordt duidelijk dat de equivalering zinvol is? Op die vraag heeft IRT een antwoord.

In de IRT staat het te meten begrip of de te meten eigenschap centraal. De IRT beschouwt het antwoord op een item als een indicator voor de mate waarin die eigenschap aanwezig is. Het verband tussen eigenschap en itemantwoord is van probabilistische aard en wordt weergegeven in de zogenaamde itemresponsfunctie. Die geeft aan hoe groot de kans is op een correct antwoord als functie van de onderliggende eigenschap of vaardigheid. Formeler: zij X_i de toevalsvariabele die het antwoord op item i voorstelt. X_i neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord.

Als symbool voor de vaardigheid wordt θ (theta) gekozen. De vaardigheid θ is niet rechtstreeks observeerbaar. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd⁸. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 | \theta) \quad (2.1)$$

Een IRT-model is een speciale toepassing van (2.1) waarbij aan de functie $f_i(\theta)$ een meer of minder specifieke functionele vorm wordt toegekend. Een eenvoudig en zeer populair voorbeeld is het zogenaamde Raschmodel (Rasch, 1960) waarin $f_i(\theta)$ gegeven is door

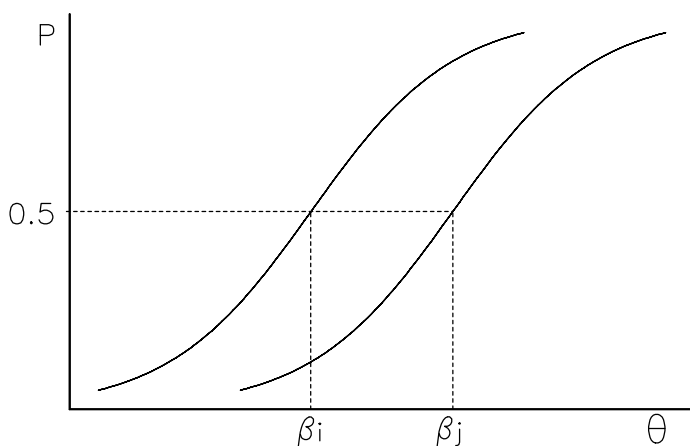
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \quad (2.2)$$

waarin β_i de moeilijkheidsparameter van item i is. Dat is een onbekende grootte die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.2 voor twee items, i en j , die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende functie is van θ : hoe groter de vaardigheid, des te groter de kans op een juist antwoord. Indien de latente vaardigheid precies gelijk is aan de moeilijkheidsparameter β_i , volgt

$$f_i(\beta_i) = \frac{\exp(\beta_i - \beta_i)}{1 + \exp(\beta_i - \beta_i)} = \frac{1}{1 + 1} = \frac{1}{2} \quad (2.3)$$

Daaruit volgt onmiddellijk een interpretatie voor de parameter β_i : het is de 'hoeveelheid' vaardigheid die nodig is voor de kans van precies een half om het item i juist te beantwoorden. Uit de figuur blijkt duidelijk dat voor item j een grotere vaardigheid nodig is om diezelfde kans te bereiken, maar dit is hetzelfde als te zeggen dat item j moeilijker is dan item i . De parameter β_i kan dus terecht omschreven worden als de moeilijkheidsparameter van item i . De implicatie van het bovenstaande is dat 'moeilijkheid' en 'vaardigheid' op dezelfde schaal liggen.

Figuur 2.2 Twee itemresponscurven in het Raschmodel



Formule (2.2) is geen beschrijving van de werkelijkheid, het is een hypothese over de werkelijkheid die getoetst kan worden op haar houdbaarheid. Hoe zo'n toetsing grofweg verloopt, is te verduidelijken aan de hand van figuur 2.2. Daaruit blijkt dat, voor welk vaardigheidsniveau dan ook, de kans om item j juist te beantwoorden steeds kleiner is dan de kans op een juist antwoord op item i . Hieruit volgt de statistisch te toetsen voorspelling dat de verwachte proportie juiste antwoorden op item j kleiner is dan op item i in een

⁸ Dit maakt duidelijk waarom men de modellen die ressorteren onder de IRT ook wel aanduidt met 'latente trek'-modellen.

willekeurige steekproef van personen. Splitst men nu een grote steekproef in bijvoorbeeld twee deelsteekproeven, een 'laaggroep', met de vijftig procent laagste scores, en een 'hooggroep', met de vijftig procent hoogste scores, dan kan men nagaan of de geobserveerde p-waarden van de opgaven in beide deelsteekproeven op dezelfde wijze geordend zijn. Daarvan kan strikt genomen alleen sprake zijn als, in termen van de klassieke testtheorie uitgedrukt, alle opgaven eenzelfde discriminatie-index hebben. Dat echter blijkt lang niet altijd zo te zijn. Ook in ons geval niet. Veel van de items blijken dan ook niet beschreven te kunnen worden met het Raschmodel. Daarom is bij de Eindtoets gekozen voor een ander IRT-model.

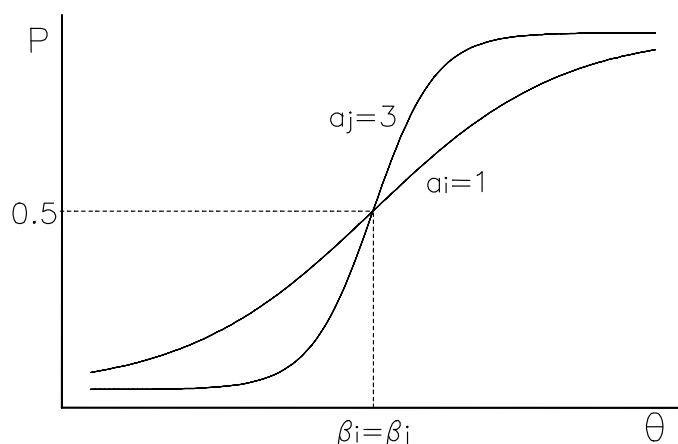
Alvorens het bij de Eindtoets gebruikte model te introduceren, is eerst een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattingsmethode is de 'conditionele grootste aannemelijkheidsmethode' (in het Engels: Conditional Maximum Likelihood, verder aangeduid als CML). Die maakt gebruik van het feit dat in het Raschmodel een afdoende steekproef-grootheid ('sufficient statistic') bestaat voor de latente variabele θ , namelijk de ruwe score of het aantal correct beantwoorde items. Dat betekent grofweg dat, indien de itemparameters bekend zijn, alle informatie die het antwoordpatroon over de vaardigheid bevat, kan worden samengevat in de ruwe score; het doet er dan verder niet meer toe welke opgaven goed en welke fout zijn gemaakt. Hieruit vloeit voort dat de conditionele kans op een juist antwoord op item i , gegeven de ruwe score, een functie is die alleen afhankelijk is van de itemparameters en onafhankelijk van de waarde van θ ⁹. De CML-schattingsmethode maakt van deze functie gebruik. Deze methode maakt geen enkele veronderstelling over de verdeling van de vaardigheid in de populatie, en is ook onafhankelijk van de wijze waarop de steekproef is getrokken.

De CML-schattingsmethode is echter niet bij elk meetmodel toepasbaar. In het zogenaamde éénparameter logistisch model (One Parameter Logistic Model, afgekort: OPLM) is CML mogelijk. Dit model is, anders dan het Raschmodel, wel bestand tegen 'omwisseling' van 'proporties juist' in verschillende steekproeven (Glas & Verhelst, 1993; Eggen, 1993; Verhelst & Kleintjes, 1993). De itemresponsfunctie van het OPLM is gegeven door

$$f_i(\theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}, \quad (2.4)$$

waarin a_i de zogenaamde discriminatie-index van het item is. Door deze indices te beperken tot (positieve) gehele getallen, en door ze a priori als constanten in te voeren, is het mogelijk CML-schattingen van de itemparameters β_i te maken. In figuur 2.3 is de itemresponscurve weergegeven van twee items i en j , die even moeilijk zijn maar verschillend discrimineren.

Figuur 2.3 Twee itemresponscurven in het OPLM: zelfde moeilijkheid, verschillende discriminatie-index



⁹ Een gedetailleerde uiteenzetting hierover kan men vinden in Verhelst, 1992.

De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas en Verstralen, 1995). Dit programma voert ook statistische toetsen uit op grond waarvan kan worden bepaald of het model de gegevens adequaat beschrijft. Omdat een aantal van deze toetsen bijzonder gevoelig is voor een verkeerde specificatie van de discriminatie-indices, zijn de uitkomsten van deze toetsen bruikbaar als modificatie-indices: ze geven een aanwijzing in welke richting deze discriminatie-indices moeten worden aangepast om een betere overeenkomst tussen model en gegevens te verkrijgen. Kalibratie van items volgens het OPLM is dan ook een iteratief proces waarin alternerend de modelfit van items wordt onderzocht door middel van statistische toetsing en de waarden van de discriminatie-indices worden aangepast op grond van de resultaten van deze toetsen. Deze aanpassingen geschieden in de praktijk op basis van een en hetzelfde gegevensbestand. Er kan dus kanskapitalisatie optreden. Indien een steekproef een voldoende grootte heeft, is het effect van deze kanskapitalisatie echter gering (Verhelst, Verstralen en Eggen, 1991).

Hoewel het OPLM aanzienlijk flexibeler is dan het Raschmodel, heeft het met dit model toch een nadeel gemeen, waardoor het bij het kalibreren van meerkeuzeopgaven niet zonder meer bruikbaar is. Uit de formules (2.2) en (2.4) volgt dat, indien θ zeer klein is, de kans op een juist antwoord zeer dicht in de buurt van nul komt. Maar de items in het normeringsonderzoek zijn meerkeuze-items, zodat blind gokken een zekere kans op een juist antwoord impliceert. Er bestaan modellen die rekening houden met de raadkans (Lord & Novick, 1968), maar die laten geen CML-schattingmethode toe. De ongeschiktheid van het Raschmodel of OPLM voor meerkeuzevragen is echter relatief: indien de items in vergelijking met de vaardigheid van de leerling niet al te moeilijk zijn, blijkt dat het effect van het raden op de overeenkomst tussen model en gegevens klein is. Door een verstandige dataverzamelingsprocedure toe te passen en met name niet te moeilijke opgaven te selecteren in de toets kan het OPLM toch toegepast worden op meerkeuzevragen, waarbij de overeenkomst tussen model en data de uiteindelijke doorslag over die geschiktheid moet geven. Ook in de normering wordt hiermee rekening gehouden.

Voor de schatting van parameters van de populatieverdeling wordt gebruikgemaakt van de 'marginale grootste aannemelijkheidsmethode' (in het Engels: Marginal Maximum Likelihood, verder afgekort als MML). Deze schattingmethode veronderstelt naast (2.2) ook nog dat de vaardigheid θ in de populatie een bepaalde verdeling heeft. De meeste computerprogramma's die IRT-analyses kunnen uitvoeren, veronderstellen een normale verdeling. Bovendien stelt deze methode de voorwaarde dat de steekproef die voor de schatting gebruikt wordt uit die verdeling een aselechte steekproef is.

3 Beschrijving van de toets

3.1 Opbouw en structuur

Inhoud en samenstelling

De Eindtoets bestaat uit twaalf taken. In de taken staan opgaven op het gebied van Taal, Rekenen-Wiskunde, Studievaardigheden en Wereldoriëntatie. Taal, Rekenen-Wiskunde en Studievaardigheden zijn verplichte onderdelen. Wereldoriëntatie is een facultatief onderdeel en speelt voor de primaire functie van de toets geen rol. Hieronder treft u een overzicht aan van de onderdelen per vakgebied (tussen haakjes staat het aantal opgaven).

Taal (totaal 100)

- Schrijven van teksten (30)
- Spelling (20)
- Begrijpend lezen (30)
- Woordenschat (20)

Rekenen-Wiskunde (totaal 60)

- Getallen en bewerkingen (25)
- Verhoudingen, breuken en procenten (20)
- Meten, meetkunde, tijd en geld (15)

Studievaardigheden (totaal 40)

- Hanteren van studieteksten (10)
- Hanteren van informatiebronnen (10)
- Lezen van schema's, tabellen en grafieken (10)
- Kaartlezen (10)

Wereldoriëntatie (totaal 90)

- Aardrijkskunde (30)
- Geschiedenis (30)
- Natuuronderwijs (30)

De verdeling van de 200 opgaven over de verschillende kennisgebieden ligt sinds jaar en dag vast. Cito heeft bewust gekozen voor deze 'gewichten' door deze een afspiegeling te laten vormen van de doorsnee onderwijsbesteding zoals deze voor het reguliere basisonderwijs bekend is. Als we bijvoorbeeld de verschillende PPON-balansen die over een reeks van jaren zijn verschenen als uitgangspunt zouden nemen, dan zouden 48% van de opgaven taalopgaven moeten zijn (overeenkomend met 7,25 uur onderwijsbesteding in groep 8) en 33% van de opgaven rekenopgaven (overeenkomend met 5 uur onderwijsbesteding in groep 8). Deze percentages komen dus vrijwel overeen met de percentages in de toets (50% en 30%). De overige opgaven betreffen studievaardigheden. 'Studievaardigheden' wordt niet als afzonderlijk vak onderwezen in het basisonderwijs, maar is onlosmakelijk verbonden met alle onderwezen vakken in het basisonderwijs. Opgaven met betrekking tot dit onderdeel komen voor 20% in de toets voor. Het vak Wereldoriëntatie kan in aanvulling op de Eindtoets afzonderlijk worden getoetst. Het betreft 90 opgaven die in gelijke verhoudingen zijn verdeeld over 'Aardrijkskunde', 'Geschiedenis' en 'Natuur'. Deze verdeling binnen Wereldoriëntatie komt overeen met de gelijke verdeling over deze vakken zoals deze in de PPON-balansen is terug te vinden (namelijk per vakgebied één uur per week in groep 8).

De opgaven zijn opgenomen in drie opgavenboekjes die op drie achtereenvolgende toetsdagen worden afgenomen. Daarbij is gezorgd voor een optimale spreiding van taken. Binnen taken is ook steeds sprake van voldoende variatie in onderdelen. Waar dat technisch nodig is zijn opgaven van een onderdeel bij elkaar gehouden. Taken op het gebied van Taal en Rekenen-Wiskunde zijn over meerdere toetsdagen gespreid, terwijl het onderdeel Studievaardigheden in zijn geheel op de tweede dag aan de orde komt. De 90 facultatieve opgaven voor Wereldoriëntatie zijn eveneens over de drie toetsdagen gespreid, met Natuuronderwijs op dag 1, Aardrijkskunde op dag 2 en Geschiedenis op dag 3.

Dag Opgavenboekje	Taak	Omschrijving onderdeel	N items
1	Taal 1	Spellen van niet-werkwoordsvormen	10
		Spellen van werkwoordsvormen	10
		Begrijpend lezen	10
	Rekenen 1	Getallen en bewerkingen	6
		Verhoudingen, breuken en procenten	7
		Metten, meetkunde, tijd en geld	7
		Schrijven van teksten	20
	[facultatief] Wereldoriëntatie 1	Natuuronderwijs	30
		<i>Totaal dag 1</i>	<i>100</i>
2	Studievaardigheden 1	Hanteren van studieteksten	10
		Hanteren van informatiebronnen	10
	Rekenen 2	Getallen en bewerkingen	14
		Verhoudingen, breuken en procenten	6
	Studievaardigheden 1	Lezen van schema's, tabellen en grafieken	10
		Kaartlezen	10
	[facultatief] Wereldoriëntatie 2	Aardrijkskunde	30
		<i>Totaal dag 2</i>	<i>90</i>
3	Taal 3	Woordenschat	20
		Schrijven van teksten	10
	Rekenen 3	Getallen en bewerkingen	5
		Verhoudingen, breuken en procenten	7
		Metten, meetkunde, tijd en geld	8
		Begrijpend lezen	20
	[facultatief] Wereldoriëntatie 3	Geschiedenis	30
		<i>Totaal dag 3</i>	<i>100</i>
	Totaal	[Exclusief Wereldoriëntatie]	200
	Totaal	[Inclusief Wereldoriëntatie]	290

3.2 Inhoudsverantwoording

In deze paragraaf wordt de inhoud van twaalf onderscheiden taken in de Eindtoets Basisonderwijs op hoofdlijnen besproken en verantwoord. Voor een uitgebreidere verantwoording wordt verwezen naar het zogenaamde Doelenboek (Staphorsius, 2009), waarin de inhoud van de Eindtoets zeer uitgebreid wordt besproken; dit Doelenboek wordt jaarlijks (zo nodig) aangepast en ter beschikking gesteld aan de deelnemende scholen. Het zou te ver voeren om hier voorbeelden van toetsopgaven op te nemen voor alle onderscheiden rubrieken. Daarvoor verwijzen we eveneens naar het Doelenboek en natuurlijk naar de opgavenboekjes zelf. Ook is ervan afgezien om voor alle opgaven aan te geven bij welke precieze categorie, leerdoel of onderdeel deze aansluiten. Belangrijkste reden daarvoor is dat de inhoud van de toets jaarlijks volledig wordt vernieuwd.

De Eindtoets Basisonderwijs is methodeonafhankelijk en gebaseerd op de kerndoelen van het basisonderwijs zoals die gegeven worden door het Ministerie van OCW. Deze kerndoelen zijn door Cito verder uitgewerkt en beschreven in het Doelenboek. Vervolgens is het Doelenboek gebruikt als 'toetsmatris' voor de Eindtoets, waardoor deze toets precies aansluit bij de kerndoelen. De combinatie van de uitgebreide verantwoording in termen van het Doelenboek en de zorgvuldige constructieprocedures

(zoals het voortdurend analyseren van de inhoud van onderwijsmethoden, het construeren van items in itembanken, het opzetten van proeftoetsen met een veelvoud aan benodigde items, het inschakelen van ervaren leerkrachten in het constructieproces) zorgt ervoor dat de toetsinhoud representatief is en blijft. We hebben ervan afgezien om een meer theoretische uiteenzetting te geven van de wijze waarop de leerling zich verschillende reken-, taal- en studievoordigheden eigen maakt. Men kan de vorderingen van de leerling nagaan door middel van (bijvoorbeeld) het leerlingvolgsysteem LOVS. In de diverse wetenschappelijke verantwoordingen die bij de verschillende LOVS-toetsen zijn verschenen worden de betreffende constructen uitgebreid theoretisch ingekaderd en wordt ruim aandacht gegeven aan de wijze waarop de leerling de betreffende vaardigheden verwerft. Een dergelijke uiteenzetting valt buiten het bestek van de Eindtoets waarbij de onderwijsdoelen (als eindpunten van het leerproces bij het afsluiten van het basisonderwijs) het uitgangspunt zijn. De geïnteresseerde lezer verwijzen we naar de genoemde wetenschappelijke verantwoordingen.

3.2.1 Taal

Het doel van het moedertaalonderwijs is het optimaliseren van de mogelijkheden van leerlingen om door gebruik van taal met anderen te communiceren. De hoofdstructuur van het taalonderwijs in het primair onderwijs sluit in grote lijnen aan bij de indeling van kerndoelen basisonderwijs (Ministerie van OCenW, 1998) en valt uiteen in de onderdelen produceren van teksten (spreken en schrijven), begrijpen van teksten (luisteren en lezen), algemene en specifieke ondersteunende vaardigheden (spreken, schrijven, luisteren, lezen). Niet alle onderdelen van dit taalonderwijs lenen zich ertoe om opgenomen te worden in een schriftelijke meerkeuzetoets. Waar dit wel mogelijk is, ligt de volgende indeling van leerdoelen ten grondslag aan de taalopgaven in de Eindtoets:

- Producenten van teksten
 - Schrijven (*'Schrijven van teksten'*)
 - Inhoud
 - juistheid van de in de tekst opgenomen informatie
 - afstemming van de informatie op het doel en de lezer
 - Taalgebruik
 - correct taalgebruik
 - interpreteerbaar taalgebruik
 - passend taalgebruik
- Ondersteunende vaardigheden voor het schrijven
 - Spelling (*'Spelling'*)
 - niet-werkwoorden
 - werkwoorden
- Begrijpen van teksten
 - Lezen: begrijpend lezen (*'Begrijpend lezen'*)
- Algemene 'ondersteunende' vaardigheden
 - Woordenschat (*'Woordenschat'*)
 - Idioom

Schrijven van teksten

De schrijfvaardigheidstaken in de Eindtoets bestaan uit teksten, geschreven door leerlingen van het basisonderwijs zelf (maar aangepast door Cito), inclusief de karakteristieke fouten die leerlingen daarbij maken. Het schrijven van teksten is hier gedefinieerd als het overdragen van informatie met een bepaald doel. Daarbij worden door de schrijver keuzes gemaakt met betrekking tot de (selectie van) inhoud en het taalgebruik in de tekst.

Bij de keuze van informatie maakt de (onervaren) schrijver fouten: onnauwkeurigheid, men houdt zich niet aan de feiten, onvoldoende of vage informatie, interne tegenspraak of andere onlogische constructies. Bij het afstemmen van de informatie op het doel en de lezer zijn de criteria voor wat fout is wat minder scherp te omschrijven. De belangrijkste fouten betreffen een overdaad aan informatie en het niet of onvoldoende rekening houden met de voorkennis van de lezer.

Verder is het taalgebruik van belang. Bij het schrijven moeten *woorden* in de juiste vorm worden geschreven, dus rekening houdend met daarover bestaande conventies. Hetzelfde geldt voor (ontsporingen in) *zinnen*: onjuiste samentrekkingen, kromme zinnen, onjuiste werkwoordtijden, incongruentie van persoonsvorm en onderwerp, het onjuiste gebruik van leestekens en hoofdletters. Al deze 'ontsporingen' kunnen de basis vormen van opgaven in dit deel van de Eindtoets. Correctheid op het niveau van de *tekst* (bijvoorbeeld voortvloeiend uit conventies over de indeling van brieven en dergelijke) komt in de Eindtoets niet aan de orde.

Verder moeten woorden, zinnen en teksten als geheel begrijpelijk zijn. Dat betekent bijvoorbeeld dat in de teksten geen dubbelzinnige formuleringen, overbodige woorden en onduidelijk of verkeerd uitgedrukte relaties tussen en binnen zinnen mogen voorkomen. In de opgaven moeten leerlingen ook de inhoudelijke structuur van een tekst of fragment daaruit beoordelen en verbeteren.

Ten slotte moet het taalgebruik ook passend zijn, dit zowel in termen van het doel van de schrijver als in termen van de relatie met zijn publiek. De woordkeus moet bijvoorbeeld afgestemd zijn op de kennis van de lezer en de toon van een tekst, zinnen mogen niet te gecompliceerd zijn en ook moeten woorden passen bij het karakter van een tekst als geheel (bijvoorbeeld een deftige uitdrukking in een verder zeer informele tekst misstaat).

Spelling

Het spellen valt uiteen in het spellen van niet-werkwoorden en het spellen van werkwoorden.

Bij *het spellen van niet-werkwoorden* gaat het vooral om de volgende categorieën van spellingproblemen:

- woorden met tweeklanken en klankcombinaties zoals *au, ou, ie, aai, ooi, eeu, ieuw, ng, nk, ing* enzovoort;
- woorden met een open lettergreep;
- woorden beginnend met *f*, *s*- of *z*-;
- woorden waarvan bij omvorming de *-s* en de *-f* respectievelijk een *-z* en *-v* worden;
- woorden met *ch, sch, schr*;
- woorden met een dubbele medeklinker;
- stoffelijke bijvoeglijke naamwoorden;
- eigennamen en daarvan afgeleide bijvoeglijke naamwoorden;
- woorden voorafgegaan door *'s*;
- woorden eindigend op *-tie*;
- woorden eindigend op *-ische*.

Bij *het spellen van werkwoorden* gaat het om een adequate toepassing van de regels van de werkwoordspelling in de volgende vormen:

- de persoonsvorm;
- de infinitief;
- het voltooid en tegenwoordig deelwoord;
- het bijvoeglijk gebruikt deelwoord.

Strikt genomen zijn de regels van de werkwoordspelling van toepassing op de persoonsvorm. De andere werkwoordsvormen zijn (didactisch gezien) een probleem omdat leerlingen die regels generaliseren naar niet-persoonsvormen.

Begrijpend lezen

De taken Begrijpend lezen bestaan uit teksten – fictie en non-fictie – met meerkeuzeopgaven over deze teksten. De opgaven komen in grote lijnen neer op een van de volgende vragen naar de betekenis van de tekst: *wat* zegt de schrijver van de tekst *waarover* met *welk doel* en *voor wie* is zijn tekst bestemd? Als de vragen betrekking hebben op grotere fragmenten staat de conceptuele betekenis of de *inhoud* van de tekst centraal. Daarnaast is voor de betekenis van de tekst ook het 'hoe' van de tekst heel belangrijk.

Toetsopgaven die daarbij aansluiten vragen naar de kenmerken van de *situatie* waarin gecommuniceerd wordt, dus naar bedoelingen van de schrijver en de relatie met zijn publiek (zie ook onder 'Schrijven van teksten'). Een tekst zegt iets over de 'situatie' door de keuzes van de schrijver met betrekking tot taalgebruik en inhoud. Ook andere aspecten van taalgebruik dan woordkeus, zoals zinsbouw (lengte, complexiteit) en stijl (persoonlijk, ambtelijk), dragen bij aan de betekenis van een tekst.

Bij Begrijpend lezen kunnen de opgaven naast de vorm van vragen bij teksten ook de vorm hebben van zogenaamde husselteksten en teksttoetsen. Deze vormen worden hieronder toegelicht.

- *Vragen bij teksten*

Bij de constructie van vragen bij teksten wordt vooral uitgegaan van de hieronder opgesomde inhoud:

- de bedoeling van de schrijver;
- het publiek van de schrijver (voor wie is de tekst bestemd?);
- typering van eigenschappen of kenmerken van de schrijver; zijn houding, uitgangspunten, waarden of ideeën;
- typering van wat de schrijver met de tekst doet (bijvoorbeeld een verslag geven, grapje maken, waarschuwen);
- de kernvraag van de tekst of het antwoord daarop (de hoofdgedachte);
- het onderscheid tussen de kern of hoofdgedachte en de daaraan ter ondersteuning toegevoegde informatie;
- een passende titel of verklaring van de gebruikte titel;
- het onderscheid tussen feit en mening;
- de bruikbaarheid van de tekst voor een gegeven leesdoel;
- de centrale gebeurtenis in een verhaal;
- de strekking of moraal van een tekst of verhaal;
- feiten en relaties daartussen;
- typering van eigenschappen of kenmerken van figuren in de tekst; hun houding, uitgangspunten, waarden of ideeën;
- typering van situaties;
- de betekenis van woorden en woordgroepen;
- de inhoudelijke structuur van de tekst;
- de bedoeling van woorden of woordgroepen die voor hun betekenis geheel of gedeeltelijk afhankelijk zijn van wat eerder in de tekst aan de orde is geweest of van wat later nog aan de orde zal komen.

- *Husselteksten*

Het uitgangspunt is een korte tekst waarvan de zinnen 'gehusseld' zijn. Van leerlingen wordt gevraagd de zinnen in de juiste volgorde te plaatsen of de eerste zin te identificeren.

- *Teksttoetsen*

Deze opgavenvorm bestaat uit een tekst waaruit woorden of zinnen zijn weggelaten. Van de leerling wordt gevraagd te midden van een aantal gegeven woorden of zinnen het woord of de zin te herkennen waarmee de tekst inhoudelijk of qua taalgebruik het beste weer compleet kan worden gemaakt.

Woordenschat (en idiomatische kennis)

In het onderdeel Taal zijn opgaven 'woordenschat' en 'idioom' opgenomen. Daarbij wordt de woordenschat opgevat als een algemene ondersteunende vaardigheid die een rol speelt bij het produceren en begrijpen van teksten. Verschillen in woordenschat blijken zeer hoog samen te hangen met verschillen in taalvaardigheid. Daarnaast kent de taak Schrijven van teksten een aantal opgaven die betrekking hebben op idiomatische kennis. Ze vragen naar het gebruik en de betekenis van vaste uitdrukkingen.

In de opgaven komt het volgende aan de orde:

- van een gegeven woord of uitdrukking de juiste betekenisomschrijving herkennen;
- een tegenstelling of synoniem van een gegeven woord aangeven;
- woorden herkennen die tot een gegeven betekenisveld behoren;
- mogelijke combinaties van een gegeven woord met andere woorden onderscheiden van onmogelijke combinaties;
- figuurlijk gebruik van woorden herkennen.

3.2.2 Rekenen-Wiskunde

In de kerndoelen basisonderwijs (Ministerie van OCW, 1998) wordt Rekenen-Wiskunde als volgt getypeerd: Het onderwijs in Rekenen-Wiskunde is erop gericht dat de leerlingen:

- verbindingen kunnen leggen tussen het onderwijs in Rekenen-Wiskunde en hun dagelijkse leefwereld;
- basisvaardigheden verwerven, eenvoudige wiskundetaal begrijpen en toepassen in praktische situaties;
- reflecteren op eigen wiskundige activiteiten en resultaten daarvan op juistheid controleren;
- eenvoudige verbanden, regels, patronen en structuren opsporen;
- onderzoeks- en redeneerstrategieën in eigen woorden kunnen beschrijven en gebruiken.

In de bestaande methoden op het gebied van Rekenen-Wiskunde worden deze kerndoelen uitgewerkt in leerlijnen, leerinhouden en leerdoelen. De leerdoelen zijn in de volgende drie domeinen in te delen:

- Getallen en bewerkingen;
- Verhoudingen, breuken en procenten;
- Meten, meetkunde, tijd en geld.

Getallen en bewerkingen

- *Basiskennis en begrip van natuurlijke getallen en decimale getallen*
 - Uitspraak en schrijfwijze van natuurlijke getallen en decimale getallen;
 - Structuur van de telrij;
 - Structuur van getallen;
 - Vergelijken en ordenen van getallen en meetuitkomsten;
 - Getallen afronden op een tiende, een honderdste, een duizendste, een geheel getal, een tiental, een honderdtal, duizendtal, tienduizendtal, honderdduizendtal, miljoen.

- *Hoofdrekenen*

Optellen, aftrekken, vermenigvuldigen en delen met getallen sec en in toepassingssituaties, waarbij gebruik wordt gemaakt van basiskennis van getallen en van inzicht in relaties tussen getallen en eigenschappen van bewerkingen. Geen gebruik van uitrekenpapier; berekeningen zonder veel geheugenbelasting.

- Optellen
Optellen met natuurlijke getallen en kommagetallen, waarbij werkwijzen gebruikt kunnen worden als hergroeperen, splitsen, doortellen met sprongen en één of meer getallen veranderen en daarvoor zo nodig een correctie toepassen.
- Aftrekken
Aftrekken met natuurlijke getallen en kommagetallen, waarbij werkwijzen gebruikt kunnen worden als hergroeperen, aanvullen, splitsen, doortellen met sprongen, twee getallen in één keer aftrekken en één of meer getallen veranderen en daarvoor zo nodig een correctie toepassen.
- Vermenigvuldigen
Vermenigvuldigen met natuurlijke getallen en kommagetallen, waarbij werkwijzen gebruikt kunnen worden als hergroeperen, splitsen of verwisselen, één of beide getallen veranderen en daarvoor zo nodig een correctie toepassen.
Vermenigvuldigen waarbij met nullen gerekend moet worden.
- Delen
Delen met natuurlijke getallen en kommagetallen, waarbij werkwijzen gebruikt kunnen worden als redeneren vanuit delen als het omgekeerde van vermenigvuldigen, splitsen, één of beide getallen veranderen en daarvoor zo nodig een correctie toepassen.
Delen waarbij met nullen gerekend moet worden.
Delen in toepassingssituaties met rest.
- Combinaties
Oplossen van opgaven waarbij verschillende operaties uitgevoerd moeten worden.

- *Schattend rekenen*

Optellen, aftrekken, vermenigvuldigen en delen met afgeronde getallen sec en in toepassingssituaties.

Geen gebruik van uitrekenpapier, berekeningen zonder veel geheugenbelasting.

Schatten via optellen, aftrekken, vermenigvuldigen of delen met afgeronde getallen dan wel in combinaties met deze bewerkingen.

- *Kolomsgewijs rekenen en cijferen*

Optellen, aftrekken, vermenigvuldigen en delen met getallen sec en in toepassingssituaties, waarbij gebruik gemaakt kan worden van uitrekenpapier, waarbij cijfernotaties en tussenuitkomsten van berekeningen ter ondersteuning van het geheugen geschreven kunnen worden.

- Optellen

Optellen met natuurlijke getallen en kommagetallen sec en in toepassingssituaties, waarbij sprake is van samenvoegen en toevoegen.

- Aftrekken

Aftrekken met natuurlijke getallen en kommagetallen sec en in toepassingssituaties, waarbij sprake is van eraf halen, splitsen, aanvullen en verschil bepalen.

- Vermenigvuldigen

Vermenigvuldigen met natuurlijke getallen en kommagetallen sec en in toepassingssituaties.

- Delen

Delen met natuurlijke getallen en kommagetallen sec en in toepassingssituaties, waarbij sprake is van verdelen en opdelen en waarbij ook de rest geïnterpreteerd moet worden.

- Combinaties

Oplossen van opgaven waarbij verschillende operaties uitgevoerd moeten worden

Oplossen van opgaven waarbij het gemiddelde berekend moet worden.

Verhoudingen, breuken en procenten

- *Verhoudingen*

Basiskennis, begrip en toepassingen

- Beoordelen of figuren in eenzelfde verhouding zijn weergegeven;
- Omzettingen van beschrijvingen met verhoudingsgetallen in beschrijvingen met een breuk of percentage of omgekeerd, zowel exact als globaal schattend;
- Oplossen van elementaire verhoudingsproblemen, waarbij of de verhouding, of een deel, of het geheel moet worden bepaald (vergrotingen, verkleiningen, ontbrekende evenredige berekenen);
- Werkelijke afstand bepalen met behulp van een schaallijn en een liniaal;
- Vergelijken van gegevens op basis van verhoudingsrelaties.

- *Breuken*

Basiskennis, begrip en toepassingen

- Beschrijven met breuken van deel-geheelrelaties en verdelingen;
- Globaal en precies plaatsen van breuken / gemengde getallen op een getallenlijn;
- Vereenvoudigen, compliceren;
- Vergelijken en ordenen van breuken / gemengde getallen;
- Omzetten van breuken naar decimale breuken en omgekeerd; zowel precies als globaal schattend;
- Oplossen van deel-geheelproblemen waarbij een deel van een aantal, bedrag of hoeveelheid berekend wordt of waarbij het totaal berekend wordt als een deel gegeven is;
- Uitvoeren van andere elementaire operaties (optellen, aftrekken, vermenigvuldigen en delen) met breuken en gemengde getallen.

- *Procenten*

Basiskennis, begrip en toepassingen

- Beschrijven van deel-geheelrelaties met procenten, zowel precies als globaal schattend;
- Percentages aflezen in een grafiek (cirkeldiagram, strook-, staafgrafiek);

- Omzetten van breuken, natuurlijke getallen, kommagetallen en verhoudingen naar procenten, zowel precies als globaal schattend;
- Oplossen van elementaire procentproblemen (zowel precies als globaal schattend) in contexten met onder andere korting, toename, afname, stijging, daling, winst, verlies, BTW, renteberekeningen. Hierbij komen zowel opgaven voor die hoofdrekenend zijn op te lossen als opgaven waarbij gebruik van uitrekenpapier bij het oplossen is toegestaan.

Metten, meetkunde, tijd en geld

- *Lengte*
 - Notie van lengtematen en het gebruik daarvan;
 - Aflezen van een meetresultaat op liniaal/meetlint;
 - Interpreteren van lengteaanduidingen op (bouw)tekeningen;
 - Interpreteren van gegevens op kilometertellers;
 - Kiezen van de juiste maat in een gegeven context;
 - Herleiden sec en in toepassingssituaties;
 - Omtrek berekenen van rechthoeken, rechthoekige figuren en benaderen van grillige figuren;
 - Oplossen van elementaire toepassingsproblemen waarbij herleidingen en berekeningen met lengtematen uitgevoerd worden.
- *Oppervlakte*
 - Notie van oppervlaktematen en het gebruik daarvan;
 - Eenvoudige herleidingen met oppervlaktematen;
 - Bepalen oppervlakte via afpassen met natuurlijke (of gegeven) maten;
 - Oplossen van oppervlakteproblemen door oppervlakte exact te bepalen (eventueel via omvormen);
 - Oplossen van oppervlakteproblemen door oppervlakte bij benadering te bepalen.
- *Inhoud*
 - Aflezen van de inhoud op een schaalverdeling, al dan niet in combinatie met een herleiding (globaal en precies);
 - Herleiden sec en in toepassingssituaties, bijvoorbeeld in situaties waarbij inhoudsaanduidingen vergeleken worden;
 - Notie van inhoudsmaten en het gebruik daarvan;
 - Bepalen van inhoud door een object als natuurlijke maat te gebruiken;
 - Berekenen van de inhoud met behulp van de formule $l \cdot b \cdot h$; zowel exact als globaal schattend.
- *Gewicht*
 - Aflezen, interpreteren van het resultaat van een weging op een weegschaal (ook digitale aanduidingen);
 - Herleiden sec en in toepassingssituaties (g/kg);
 - Toepassingen, onder andere berekenen van de totaalprijs als het (af te lezen) gewicht en de prijs per kg gegeven zijn; zowel exact als globaal schattend.
- *Tijd: klok en kalender*
 - Aflezen van analoge tijden en digitale tijden;
 - Omzetten van analoge tijdsaanduidingen in digitale tijdsaanduidingen en omgekeerd;
 - Bepalen van tijdsduur en tijdstip;
 - Herleidingen uitvoeren met tijdmaten (uur, kwartier, minuut, seconde);
 - Aflezen en interpreteren van gegevens van de kalender (dagen, weken, weeknummers, maanden, kwartalen, jaren, eeuwen);
 - Aflezen en interpreteren van datumaanduiding en kwartaalaanduiding;
 - Herleidingen uitvoeren met tijdmaten (dag, week, maand, kwartaal, jaar, eeuw).

- *Geld*
 - Totale waarde bepalen van munten / bankbiljetten;
 - Omwisselen van biljetten of munten in andere munten of biljetten;
 - Gepast samenstellen van bedragen met munten of biljetten;
 - Bepalen wat men terugkrijgt (eventueel ook bijbetalen om terugkrijgen te vergemakkelijken);
 - Omrekenen van bedragen naar andere valuta.
- *Combinaties*
 - Toepassingsopgaven, waarbij twee of meer inhoudelijke gebieden betrokken zijn, bijvoorbeeld omtrek en oppervlakte, lengte en inhoud, oppervlakte en prijs per eenheid;
 - Bepalen of in een situatie omtrek-, oppervlakte- of inhoudsberekening nodig is;
 - Bepalen welke gegevens nodig of overbodig zijn;
 - Schatten van onder andere oppervlakte, inhoud, lengte of gewicht op basis van ervaringsfeiten.
- *Meetkunde*
 - Oplossen van vraagstukken waarbij de ruimte meetkundig geordend en beschreven is of waarbij ruimtelijke redeneringen gemaakt moeten worden. Voorbeelden zijn:
 - het mentaal innemen van standpunten;
 - het interpreteren van plattegronden en bouwtekeningen door ruimtelijk te redeneren;
 - het herkennen van aanzichten van ruimtelijke objecten;
 - het omvormen van figuren en het in gedachten (re)construeren van bouwplaten.

3.2.3 Studievaardigheden

Het doel van het onderdeel studievaardigheden in het onderwijs kan globaal worden aangeduid met het leren

- hanteren van informatiebronnen;
- lezen van schema's, tabellen, grafieken en kaarten;
- hanteren van studieteksten.

Bij het leren hanteren van informatiebronnen staan de selectie van de bron en de zoekstrategie in de bron centraal. Het gaat hierbij om papieren en digitale informatiebronnen. Bij schema's, tabellen, grafieken en kaarten gaat het om de interpretatie van de informatie waarbij gebruikgemaakt wordt van de specifieke kenmerken van de informatiedrager. Bij studieteksten gaat het om de omzetting van informatie in de teksten in een vorm die gemakkelijk toegankelijk is en die als geheugensteun kan dienen.

Hanteren van informatiebronnen

- *Algemene vaardigheden bij het hanteren van informatiebronnen*
 - Alfabetiseren en de toepassing daarvan in allerlei contexten, zoals woordenboek, encyclopedie, telefoongids en internet.
 - Kiezen van een geschikte informatiebron en van een geschikte toegang tot die bron (boeken, tijdschriften, kranten, folders, radio- en tv-gidsen, reisgidsen, telefoongidsen, encyclopedieën, woordenboeken, internet).
 - Kiezen van adequate trefwoorden en omschrijvingen bij het zoeken naar informatie.
- *Specifieke naslagwerken*
 - Woordenboek
Kunnen opzoeken van:
 - de betekenis van woorden, spreekwoorden, zegswijzen en afkortingen;
 - de uitspraak: de lettergreep waarop de klemtoon valt;
 - de woordvormen: meervoudsvormen en verkleinwoorden;
 - de spelling.

- Encyclopedie
Kunnen aangeven onder welk trefwoord of in welk deel van de encyclopedie het meest waarschijnlijk de gevraagde informatie te vinden is.
- Telefoongids
Kunnen:
 - hanteren van de alfabetische ordening van abonnees op achtereenvolgens woonplaats, achternaam, straatnaam;
 - opzoeken van abonnee-, netnummer en/of postcode als voldoende andere gegevens bekend zijn (bijvoorbeeld woonplaats, adres en/of beroep);
 - opzoeken van bijzondere telefoonnummers (bijvoorbeeld 06-nummers en internationale toegangsnummers).
- Overige naslagwerken
Gebruiken van andere dan bovengenoemde informatiebronnen, zoals internet, bedrijvengidsen, gemeentegidsen, reisgidsen, folders en brochures. Wat betreft internet gaat het om het gebruiken van de voornaamste besturingselementen op webpagina's, zoals zoekvensters, menubalken en hyperlinks.

Kaartlezen

Leerlingen kunnen verschillende soorten kaarten lezen. Deze kaarten verschillen in schaalniveau (plattegrond van een huis, kaart van een stad, land of werelddeel) en naar inhoud (overzichtskaarten met informatie over de ligging van plaatsen, wegen en rivieren; thematische kaarten). Daarbij kan onderscheid gemaakt worden naar de deelvaardigheden: selecteren, identificeren, analyseren en interpreteren.

- *Selecteren*
Leerlingen kunnen uit verschillende kaarten, bijvoorbeeld in een atlas, de meest geschikte kaart kiezen.
- *Identificeren*
Leerlingen kunnen gegevens op de kaart opzoeken met behulp van noordpijl of windroos (richting), schaallijn, vakkenstelsel met bijbehorend register en legenda. Leerlingen kunnen hierbij de symbolen op de kaart vertalen naar een concreet beeld van de werkelijkheid: zij kunnen de kaartgegevens visualiseren. Bijvoorbeeld: op de plattegrond van een stad kunnen aangeven waar de foto van een brug genomen is.
- *Analyseren*
Een patroon in de gegevens op de kaart ontdekken door overeenkomsten en verschillen tussen gegevens op de kaart na te gaan. Bijvoorbeeld door gegevens over neerslag te combineren met gegevens over de spreiding van de bevolking, of de spreiding van hotels in één deelgebied op de kaart te vergelijken met de spreiding van hotels in een ander deelgebied.
- *Interpreteren*
Het patroon in de gegevens op de kaart verklaren met algemene kennis van de wereld. Bijvoorbeeld uit de gegevens op een hoogtekarte van de Alpenlanden de spreiding van de wintersportplaatsen verklaren. Leerlingen kunnen zich hierbij een voorstelling maken van het landschap dat uit het kaartbeeld naar voren komt: zij kunnen het kaartbeeld visualiseren.

Lezen van schema's, tabellen en grafieken

Leerlingen kunnen verschillende soorten schema's, tabellen en grafieken lezen. Daarbij is onderscheid te maken naar structuurschema's, stroomdiagrammen, enkelvoudige tabellen, kruistabellen, beeld-, staaf-, lijn- en cirkelgrafieken. Ook hier zijn weer deelvaardigheden te onderscheiden: selecteren, identificeren, analyseren en interpreteren.

- *Selecteren*
Beoordelen of een gegeven schema, tabel of grafiek geschikt is als informatiebron bij een bepaalde vraag.
- *Identificeren*
Kunnen opzoeken van gegevens in een schema, tabel of grafiek (beeldgrafiek, staafgrafiek, lijngrafiek, cirkelgrafiek), door gebruik te maken van de titel, symboolverklaringen, celgegevens, rij-opschriften, as-opschriften, schaalverdelingen, meetpunten, relatieve verdelingen en totalen.
- *Analyseren*
Een patroon ontdekken in de gegevens van één of meer tabellen en grafieken. Een patroon kan bestaan uit een trend, of uit een bepaalde combinatie van gegevens die optimaal voldoet aan een criterium.
- *Interpreteren*
Een patroon in de tabel of de grafiek verklaren met behulp van algemene kennis van de wereld.

Hanteren van studieteksten

Het hanteren van studieteksten kan het best worden beschouwd als een aspect van het leesonderwijs. In de opgaven 'Studieteksten' staat centraal:

- de omzetting van teksten in een vorm die het opnemen en verwerken van de informatie in de tekst ondersteunt en die toegankelijk is voor raadpleging als 'extern geheugen' op een later tijdstip (transformatie);
- het kritisch lezen van studieteksten.

Deze aspecten kunnen als volgt worden uitgewerkt:

- *Transformeren van studieteksten*
Er valt onderscheid te maken naar twee soorten transformaties: verbale en schematische.
Het gaat bij *verbale transformaties* om parafrase van teksten die gebaseerd zijn op de identificatie van hoofd- en bijzaken en de structuur (grote lijn) van de tekst of van tekstfragmenten. Voorbeelden van verbale transformaties zijn:
 - samenvattingen;
 - uittreksels;
 - aantekeningen;
 - hoofdonderwerp en hoofdgedachte of kern van fragmenten;
 - centrale vragen.
 Bij *schematische transformaties* gaat het om het omzetten van studieteksten in overzichtelijke figuren. Daarbij gaat het om het onderscheiden van hoofd- en bijzaken en om het herkennen van de grote lijn in de tekst.
Voorbeelden van schema's die in de opgaven voorkomen zijn:
 - tabellen;
 - stroomdiagrammen;
 - boomdiagrammen;
 - grafieken.
 In de opgaven gaat het om het herkennen van – gelet op het gebruiksdoel en de inhoud – adequate parafrases en figuren. Daarnaast zijn er opgaven waarin markerings- en oorspronkelijke teksten zijn aangebracht op hun relevantie moeten worden beoordeeld.
- *Kritisch lezen van studieteksten*
Opgaven die een concretisering zijn van 'kritisch lezen' vragen bijvoorbeeld om:
 - de beoordeling van de bruikbaarheid voor een doel;
 - het onderscheiden van feiten en meningen versus opvattingen;
 - de houdbaarheid van conclusies.

3.2.4 Wereldoriëntatie

Op basis van de kennisgebieden die genoemd staan in artikel 9.2a tot en met 9.2e van de Wet op het primair onderwijs is onder wereldoriëntatie te verstaan:

- aardrijkskunde;
- geschiedenis;
- de natuur, waaronder biologie;
- maatschappelijke verhoudingen, waaronder staatsinrichting;
- geestelijke stromingen.

De algemene doelen voor wereldoriëntatie (zie hieronder) zijn gericht op vaardigheden die nodig zijn om onder eigen verantwoordelijkheid te leren denken en handelen. Kennis wordt met het oog hierop geselecteerd, rekening houdend met de mogelijkheden van het basisschoolleerling dat vooral gericht is op het concreet waarneembare.

Algemeen doel 1

Leerlingen verwerven bij wereldoriëntatie kennis in de vorm van basisinzichten. Deze inzichten stellen hen in staat met begrip aan de samenleving deel te nemen. Het gaat hierbij zowel om (hoofdzaken in) kennis over de natuur, zoals krachten, energie, planten en dieren, als om (hoofdzaken in) kennis over menselijk handelen en de resultaten daarvan, zoals producten, omgangsvormen en idealen.

Algemeen doel 2

Leerlingen kunnen zich een beeld vormen van de manier waarop kennis verworven wordt. Zo moeten ze de waarde van bronnen als basis voor historische kennis kunnen aangeven en kunnen ze op een eenvoudige manier empirische gegevens verzamelen over een gebied. Ook moeten ze een eenvoudig onderzoekje kunnen uitvoeren voor het verkrijgen van kennis over de natuur.

Algemeen doel 3

Leerlingen kunnen informatiebronnen gebruiken als hulpmiddel bij het verwerven van kennis. Ze moeten kunnen omgaan met naslagwerken, kaarten, en met schema's, tabellen en grafieken.

Deze algemene doelen worden hieronder voor de onderscheiden kennisgebieden uitgewerkt, waarbij het waarderend en oordelend omgaan met verschijnselen (helaas) vanwege de toetsing in meerkeuzevorm buiten beschouwing moet blijven. De uitwerking van de doelen is ontleend aan de zogenoemde Domeinbeschrijvingen Wereldoriëntatie. Deze zijn tot stand gekomen in een samenwerkingsverband waarin vakinhoudelijke, ontwikkelingspsychologische en pedagogische deskundigen hebben samengewerkt. Deze deskundigen waren verbonden aan Cito, het Instituut voor Leerplanontwikkeling, de Inspectie van het Onderwijs, een universiteit of een schoolbegeleidingsdienst. Daarbij is gewerkt volgens de methode van de cultuurpedagogische discussie (zie voor een nadere toelichting het Doelenboek). Alle wereldoriëntatie-onderwerpen die in basisschoolmethoden aan de orde komen en in de kerndoelen worden genoemd, zijn vanuit de gezichtspunten van deze methode bediscussieerd. Hiertoe zijn in totaal meer dan vijftig discussies gevoerd waarbij per keer werd deelgenomen door zo'n acht tot twaalf personen.

Deze domeinbeschrijvingen zijn sinds 2002 bij Cito verkrijgbaar. Alleen stof die in deze domeinbeschrijvingen is opgenomen kan in de Eindtoets aan de orde komen. Niet alle wenselijke stof uit de Domeinbeschrijvingen komt echter op alle scholen aan de orde. Daarom worden alleen opgaven geconstrueerd op basis van stof uit de Domeinbeschrijvingen die tot de algemeen aanvaarde en gangbare leerstof behoort.

Zowel in de Domeinbeschrijvingen als in de Eindtoets is de kennis geordend in drie domeinen: aardrijkskunde, geschiedenis en natuuronderwijs. Deze ordening sluit het beste aan bij de manier waarop wereldoriëntatie op de scholen in Nederland gegeven wordt. Alle kennisgebieden die in de wet en in de kerndoelen worden genoemd als afzonderlijke leer- en vormingsgebieden zijn in deze drie kennisgebieden geïntegreerd.

Aardrijkskunde

Aardrijkskundeonderwijs is gericht op de vorming van een eigentijds beeld van de wereld. Om dat te bereiken schenkt het schoolvak aandacht aan:

- begrippen waarmee kenmerken van de bevolking en het landschap waarin ze woont kunnen worden beschreven en verklaard (bijvoorbeeld: stad, tuinbouw, migratie);
 - spreiding van verschijnselen (bijvoorbeeld: woestijnen in Afrika);
 - samenhang tussen verschijnselen in een bepaalde regio (bijvoorbeeld: samenhang tussen reliëf en bewoning in de alpenlanden);
 - de wijze waarop informatie over regio's wordt verzameld. Hierbij spelen kaarten een belangrijke rol.
- Deze aspecten worden in het aardrijkskundeonderwijs op de basisschool aangeboden in een regionale context. Hedendaagse verschijnselen in een stad, land of andere regio staan centraal.

Elke opgave heeft zowel een inhouds- als een vaardigheidsdoel. Het inhoudsdoel geeft aan over welk onderdeel van aardrijkskunde de opgave gaat. Daarbij worden zes inhoudsdoelen onderscheiden, te weten: topografie, landschappen, bevolking, landbouw, industrie en dienstverlening. Het vaardigheidsdoel geeft aan op welke vaardigheid van de leerling een beroep wordt gedaan. Er worden drie vaardigheidsdoelen onderscheiden (zie hieronder). Er zijn dus in totaal achttien mogelijke combinaties van inhoud en vaardigheid waarop de opgaven betrekking kunnen hebben.

- *Inhouden*
 - Topografie
De belangrijkste namen op de kaart van
 - Nederland;
 - (de rest van) Europa;
 - (de rest van) de wereld.
 - Landschappen
De aarde, haar landschappen en de invloed daarvan op de mens:
 - aarde en heelal;
 - de verdeling van werelddelen, oceanen en zeeën op aarde;
 - klimaten en plantengroei op aarde;
 - ontstaan van cultuurlandschappen op aarde;
 - ontstaan en bedreiging van landschappen in Nederland.
 - Bevolking
Kenmerken en spreiding van de bevolking en veranderingen daarin door migratie:
 - omvang en spreiding van de bevolking in Nederland en de wereld;
 - culturele diversiteit en migratie in Nederland, Europa en de wereld.
 - Landbouw
De werkwijze van landbouwers en hun afhankelijkheid van de omgeving:
 - productie in landbouwbedrijven;
 - ruimtelijke aspecten van landbouw (zoals de invloed van klimaat en grond);
 - invloed van landbouw op natuur en milieu.
 - Industrie
De productielijn in fabrieken en de relatie met de omgeving:
 - productie in fabrieken;
 - ruimtelijke aspecten van industrie (zoals de vestigingsplaats);
 - invloed van industrie op natuur en milieu.
 - Dienstverlening
Het belang van dienstverlening voor mensen en bedrijven:
 - verscheidenheid van dienstverlening en de relatie met welvaart;
 - de ruimtelijke aspecten van dienstverlening (zoals bereikbaarheid van winkels);
 - invloed van dienstverlening op natuur en milieu.

- **Vaardigheden**

De leerlingen kunnen

- A samenhang tussen verschijnselen in een gebied beschrijven en verklaren;
- B de ruimtelijke spreiding van verschijnselen aangeven;
- C eenvoudige methoden hanteren om informatie over een gebied te verzamelen.

Geschiedenis

Geschiedenisonderwijs benadert verschijnselen vooral vanuit hun ontstaan in de tijd. Het geeft daarbij een beeld van:

- de samenhang van verschijnselen in een bepaalde tijd in een bepaald gebied (bijvoorbeeld: het leven in de 17^e eeuw in Nederland);
- de ontwikkeling van verschijnselen in de tijd (bijvoorbeeld: de ontwikkeling van de scheepvaart door de eeuwen heen);
- de manier waarop kennis over het verleden verkregen kan worden.

Elke opgave heeft zowel een inhoudsdoel als een vaardigheidsdoel. Het inhoudsdoel geeft aan over welke periode van de geschiedenis de opgave gaat. Er zijn in het geschiedenisonderwijs tien inhoudsdoelen te onderscheiden en vier vaardigheidsdoelen. Dit leidt tot veertig mogelijk combinaties van inhoud en vaardigheid waarop de opgaven betrekking kunnen hebben.

	TIJDVAK	JAREN/PERIODEN	KENMERKENDE ASPECTEN
1	Jagers en boeren	Prehistorie	– levenswijze van jagers-verzamelaars; – ontstaan van landbouw en landbouwsamenlevingen.
2	Grieken en Romeinen	Tot 500 n.C. Oudheid	– verspreiding van de Grieks-Romeinse cultuur en de confrontatie met de Germaanse cultuur; – christendom in het Romeinse Rijk van verboden tot enig toegestane godsdienst.
3	Monniken en ridders	500 n.C. – 1000 n.C. Middeleeuwen (vroeg)	– verspreiding van het christendom in de Lage Landen; – hofstelsel en horigheid.
4	Steden en staten	1000 – 1500 Middeleeuwen (hoge en late)	– opkomst van handel en ontstaan van steden; – opkomst van de stedelijke burgerij en toenemende zelfstandigheid van steden.
5	Ontdekkers en hervormers	1500 – 1600 Vroegmoderne Tijd Nieuwe Tijd (zestiende eeuw)	– begin van de Europese overzeese expansie; – de opstand en het ontstaan van een onafhankelijke Nederlandse staat.
6	Regenten en vorsten	1600 – 1700 Vroegmoderne Tijd Nieuwe Tijd (zeventiende eeuw)	– ontstaan van handelskapitalisme en het begin van een wereldeconomie; – burgerlijk bestuur en stedelijke cultuur in Nederland.
7	Pruiken en revoluties	1700 – 1800 Vroegmoderne Tijd Nieuwe Tijd (achttiende eeuw)	– slavenarbeid op plantages en de opkomst van het abolitionisme; – streven naar grondrechten en politieke invloed van de burgerij in de Franse en Bataafse Revolutie.
8	Burgers en stoommachines	1800 – 1900 Moderne Tijd Nieuwste Tijd (negentiende eeuw)	– ontstaan van een parlementair stelsel en de toename van de volksinvloed; – industriële revolutie en de opkomst van emancipatiebewegingen.
9	Wereldoorlogen	1900 – 1950 Moderne Tijd Nieuwste Tijd (1 ^e helft 20 ^e eeuw)	– massa en mondialisering; – economische wereldcrisis; – Duitse bezetting van Nederland en de jodenvervolgung.
10	Televisie en computer	1950 – heden Moderne Tijd Nieuwste Tijd (2 ^e helft 20 ^e eeuw)	– blokvorming tussen Oost en West in de Koude Oorlog; – dekolonisatie; – sociaal-culturele veranderingen en toenemende pluriformiteit vanaf de jaren zestig.

App.	Burgerschap en sociale integratie	Staatsinrichting en geestelijke stromingen	– staatsinrichting van Nederland: rechten en plichten, rijk, provincies en gemeenten, politie en justitie; – overige organisaties; – gezin, wonen, levensstijl en vrije tijd; – migratie en multiculturaliteit; – geestelijke stromingen in de multiculturele samenleving in Nederland.
------	-----------------------------------	--	---

Vaardigheden

De leerlingen kunnen:

- A aangeven hoe de mensen in het verleden leefden;
- B verschijnselen, gebeurtenissen en personen in de tijd plaatsen, met behulp van aanduidingen van belangrijke historische perioden (bijvoorbeeld: Gouden Eeuw of middeleeuwen), van jaartallen en van tijdlijnen;
- C de ruimtelijke spreiding aangeven van verschijnselen uit het verleden;
- D het belang aangeven van bronnen met aandacht voor:
 - het terugvoeren van historische kennis naar bronnen;
 - het afleiden van kennis uit historische bronnen;
 - het herkennen van het standpunt van de schrijver van een bron in gevallen waarin een persoonlijke stellingname duidelijk naar voren komt.

Natuuronderwijs

Natuuronderwijs belicht verschijnselen uit de natuur vooral vanuit hun onderlinge samenhang. Hierbij spelen duurzaamheid en gezondheid een belangrijke rol.

Natuuronderwijs geeft een beeld van:

- diversiteit in de natuur;
- samenhang in de natuur, de plaats van de mens in en de invloed van de mens op de natuur;
- eigenschappen die organismen hebben om zich in een gebied te handhaven en gezond te blijven;
- het belang van voortplanting om leven in stand te houden;
- natuurkundige en technische principes in het dagelijks leven;
- mogelijkheden om verschijnselen uit natuur en techniek met eenvoudige hulpmiddelen te onderzoeken.

Elke opgave heeft zowel een inhoudsdoel als een vaardigheidsdoel. Het inhoudsdoel geeft aan over welk onderdeel van natuuronderwijs de opgave gaat. In het natuuronderwijs zijn vier inhoudsdoelen te onderscheiden. De eerste drie inhoudsdoelen gaan over de levende natuur. Het laatste inhoudsdoel gaat over natuurkunde en techniek. Het vaardigheidsdoel geeft aan op welke vaardigheid van de leerling een beroep wordt gedaan. Er worden vier vaardigheidsdoelen onderscheiden.

Met de vier inhoudsdoelen en de vier vaardigheidsdoelen ontstaan zestien mogelijke combinaties waarop de opgaven betrekking kunnen hebben.

- *Inhouden*
 - Organismen, waarnemen en reageren
Eenvoudige systematiek en eigenschappen:
 - globale indeling van organismen;
 - voortbeweging in het water, op het land of in de lucht;
 - eigenschappen om uitdroging op het land te voorkomen; lichaamstemperatuur;
 - waarnemen en reageren bij planten, dieren en mensen (zoals zintuigen, zenuwen, geraamte en spieren).
 - Stofwisseling en kringloop
Samenhang tussen organismen onderling en tussen organismen en hun omgeving:
 - functies van bouwstoffen, brandstoffen en vitamines:
 - voeding bij mensen, dieren en planten (zoals spijsvertering, transport, uitscheiding, opslag en het maken van bouwstoffen, brandstoffen en vitamines door planten);
 - verwerken van resten van organismen door schimmels en bacteriën.

- kringloop van gassen, voedselketens en kringloop van mineralen:
 - eigenschappen van organismen (bouw, gedrag) om zich te handhaven in het milieu (zoals door ademhaling, voeding, bescherming);
 - afstemming van processen op elkaar (zoals toename ademhaling, hartslag, zweten bij sporten; ziekte en gezondheid);
 - inwendig evenwicht van organismen en de mogelijkheid van mensen om dit evenwicht te beïnvloeden (zoals voedselkeuze, gezondheidszorg);
 - evenwicht in de natuur als geheel en de invloed van de mens daarop (zoals milieuvervuiling, biologische bestrijding van plaagdieren, natuurbehoud, biotechnologie).
- Voortplanting en ontwikkeling
 - voortplanting en ontwikkeling bij dieren, mensen en planten;
 - voortplantingscellen, voortplanting zonder bevruchting (zoals met bollen of knollen), voortplanting met bevruchting;
 - eigenschappen van planten, mensen en dieren (bouw, gedrag) waardoor bevruchting mogelijk wordt (zoals bestuiving, puberteit, territorium- en baltsgedrag):
 - belang van aanleg en milieu bij de ontwikkeling van organismen;
 - ontwikkeling bij dieren en de verzorging van eieren en/of jongen;
 - ontwikkeling bij mensen (zoals groei voor en na de geboorte);
 - ontwikkeling bij planten (zoals zaden, vruchten, zaadverspreiding en kieming).
- Natuurkunde en techniek
 Natuurkundige en technische principes bij verschijnselen in het dagelijks leven:
 - energie (energiebronnen, energieomzetting, verbranden, warmtegeleiding en aggregatietoestanden);
 - licht en kleur (lichtbron, regenboog, kleurenfilter, schaduw);
 - elektriciteit (opwekken van elektriciteit, stroomkring, stroomgeleiding, kortsluiting);
 - zinken, zweven en drijven (zwaartekracht en opwaartse kracht);
 - lucht (zuurstof, luchtdruk, vacuüm, wind);
 - geluid (geluidstrilling, geluidssnelheid, toonhoogte, geluidssterkte);
 - magnetisme (magnetische kracht, polen, kompas);
 - constructies, overbrengingen en besturingen (krachten, stevige constructies, verbindingen, tandwielen, hefboom, besturingssysteem).
- *Vaardigheden*
 De leerlingen kunnen:
 - A het belang aangeven van eigenschappen van organismen om in leven te blijven;
 - B samenhang aangeven tussen organismen onderling en tussen organismen en hun omgeving;
 - C natuurkundige en technische principes bij verschijnselen in het dagelijks leven gebruiken;
 - D verschijnselen uit natuur en techniek met eenvoudige experimenten onderzoeken.

3.3 Kenmerken van de toets: beschrijvende gegevens

Itemkenmerken: moeilijkheidsgraad en interne consistentie

Wat de moeilijkheid van de opgaven betreft: voor de opgavenselectie geldt het uitgangspunt dat de p-waarden bij voorkeur tussen .40 - .90 moeten liggen en dat de opgaven in de Eindtoets gemiddeld een p-waarde hebben van ongeveer .70 of iets hoger. In tabel 3.1 rapporteren we de range van p-waarden en de gemiddelde p-waarde van de opgaven voor de Eindtoets 2010. Ter vergelijking zijn dezelfde gegevens voor 2009 in de tabel opgenomen. De cijfers hebben betrekking op de toetsonderdelen zoals deze worden gepresenteerd in het leerlingrapport, waarbij het onderdeel Taal is opgesplitst in deelvaardigheden, dus op Rekenen-Wiskunde, Begrijpend lezen, Schrijven, Spelling, Woordenschat en Studievaardigheden. Ook worden gegevens gepresenteerd voor de facultatieve onderdelen op het terrein van Wereldoriëntatie (Aardrijkskunde, Geschiedenis en Natuuronderwijs). Daarnaast zijn ook gegevens opgenomen over de Rit-waarden van de opgaven, waarbij de toetsscore over het betreffende onderdeel het uitgangspunt was voor

de berekening van de coëfficiënt. Rit-waarden zijn wellicht te prefereren omdat zij een realistischer beeld geven van de correlatie met de schaalscore, maar helaas zijn ons geen normgegevens bekend voor Rit. Voor Rit-waarden kent het COTAN-beoordelingssysteem (Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K., 2010) wel kwaliteitscriteria. De gegevens hebben betrekking op alle leerlingen die in de betreffende leerjaren hebben deelgenomen aan de reguliere papieren versie van de Eindtoets (2009: N=138 400; 2010: N=137 301). De afname van de facultatieve onderdelen Wereldoriëntatie betreft minder leerlingen (120 147 in 2009, 118 006 in 2010).

Tabel 3.1 Range en gemiddelde van p- en Rit-waarden voor 2009 en 2010 naar toetsonderdeel

Toetsonderdeel	2009				2010			
	P-waarden		Rit-waarden		P-waarden		Rit-waarden	
	Range	Gem.	Range	Gem.	Range	Gem.	Range	Gem.
Rekenen-Wiskunde	.46 - .90	.71	.18 - .56	.40	.46 - .92	.71	.24 - .59	.42
Begrijpend lezen	.29 - .98	.77	.10 - .52	.35	.52 - .93	.77	.26 - .55	.39
Schrijven	.46 - .97	.77	.13 - .45	.35	.55 - .99	.79	.20 - .50	.38
Spelling	.55 - .97	.74	.20 - .49	.40	.51 - .96	.76	.30 - .46	.38
Woordenschat	.50 - .90	.71	.27 - .47	.37	.52 - .91	.74	.25 - .54	.39
Studievaardigheden	.54 - .94	.75	.19 - .54	.36	.49 - .95	.78	.16 - .46	.35
Aardrijkskunde	.47 - .91	.68	.23 - .52	.38	.42 - .92	.74	.19 - .51	.36
Geschiedenis	.32 - .93	.71	.22 - .46	.36	.52 - .93	.76	.24 - .46	.39
Natuuronderwijs	.35 - .93	.73	.20 - .49	.35	.41 - .92	.71	.18 - .45	.35

Wanneer we ons beperken tot 2010 blijken de p-waarden van de onderdelen te variëren tussen .71 en .79. Voor de minima per onderdeel geldt dat geen enkele laagste p-waarde onder de .40 uitkomt. De maximale p-waarden per onderdeel komen alle boven de .90 uit, maar het betreft uitzonderingen, steeds één of enkele items per onderdeel. Het COTAN-beoordelingssysteem geeft voor Rit-waarden als kwaliteitscriterium aan, dat waarden boven de .30 als goed beoordeeld worden. De gemiddelde Rit-waarden liggen tussen de .35 en .41 en voldoen dus ruim aan deze norm. Rit-waarden onder de .20 worden volgens hetzelfde criterium als onvoldoende beoordeeld. Van alle 290 opgaven zijn er slechts drie die op deze manier onder de maat zijn gebleven.

De gegevens voor 2009 zijn ter vergelijking eveneens in de tabel opgenomen. Te zien is dat beide versies op vergelijkbare wijze aan criteria voor moeilijkheidsgraad en interne consistentie voldoen. De versie van 2009 kan misschien als iets moeilijker worden gekwalificeerd dan die van 2010. De Rit-waarden zijn vergelijkbaar met gemiddelden tussen .35 en .40. In 2009 hadden (slechts) vier van de 290 items een Rit-waarde onder de .20.

Aantal goed (ruwe score) en standaardscore: scoreverdeling

De Eindtoets Basisonderwijs wordt (via equivalering; zie hoofdstuk 2) zo gemaakt dat het voor de behaalde standaardscore niet uitmaakt welke jaargang van de Eindtoets door de leerling wordt gemaakt. Bij de omzetting van het totaal aantal goed op de onderdelen Taal, Rekenen-Wiskunde en Studievaardigheden (200 opgaven) naar een standaardscore wordt rekening gehouden met de moeilijkheid van de toets. Die moeilijkheid is af te lezen aan de gemiddelde ruwe score (gemiddeld aantal goed) op de Eindtoets. In 2009 was de gemiddelde score 147.90. In tabel 3.2 is te zien dat in 2010 de Eindtoets dus gemakkelijker is geweest: in 2010 hadden de leerlingen gemiddeld 150.20 opgaven goed.

In de tabel zijn ook verdelingsgegevens op andere aggregatieniveaus dan de totale ruwe score opgenomen. De gemiddelden komen uiteraard overeen met wat men bij een gegeven aantal items mag verwachten bij de gekozen (gemiddelde) moeilijkheidsgraad. Omdat deze gemiddelde moeilijkheidsgraad voor alle onderdelen boven de .70 ligt, zijn alle verdelingen linksscheef, de ene wat meer dan de andere. Qua scheefheid ontlopen de verdelingen elkaar niet veel.

De standaardscore is een getal op een schaal van 501 tot en met 550. Het is een lineaire transformatie van het totaal aantal goed (de ruwe score) op de onderdelen Taal, Rekenen-Wiskunde en Studievaardigheden (200 opgaven) naar een schaal waarmee de 'aantallen goed' op verschillende jaargangen van de Eindtoets te vergelijken zijn. Dit houdt in dat de ruwe score wordt vermenigvuldigd met een getal (A^*) en dat er daarna een vast getal (B^*) bij wordt opgeteld. De waarden van A^* en B^* worden jaarlijks bepaald door middel van een IRT-analyse die de gegevens levert voor een daarop volgend equivaleringsonderzoek. Die equivalering levert vervolgens de A^* en B^* coëfficiënten (zie verder hoofdstuk 2).

Tabel 3.2 Psychometrische kenmerken van de Eindtoets 2010 (N=137 301); inclusief Wereldoriëntatie (N=118 006)

Onderdeel	Aantal opgaven	Gemiddelde	Standaard deviatie	Skewness	Kurtosis
Totaal	200	150.20	26.05	-.635	-.044
Taal	100	76.82	12.67	-.699	.162
Rekenen-Wiskunde	60	42.37	10.88	-.599	-.353
Studievaardigheden	40	31.01	5.62	-.865	.505
Taal					
Begrijpend lezen	30	23.19	4.63	-.865	.372
Schrijven	30	23.57	4.50	-.933	.594
Spelling	20	15.19	3.07	-.594	.068
Woordenschat	20	14.87	3.31	-.641	-.028
Aardrijkskunde	30	22.11	4.52	-.732	.268
Geschiedenis	30	22.77	4.78	-.799	.327
Natuuronderwijs	30	21.26	4.55	-.503	-.147

De verdeling van de standaardscores is gegeven in tabel 3.3. Daarbij is voor zowel 2009 als 2010 aangegeven welke ruwe scores horen bij welke standaardscore. Daarnaast zijn de cumulatieve relatieve frequenties vermeld. Deze frequentieverdeling van de scores laat zien welk 'aantal goed' er nodig is om een bepaalde standaardscore te krijgen. Voor een standaardscore van 535 moesten de leerlingen bijvoorbeeld in 2009 een aantal goed hebben van 146 t/m 148; in 2010 krijgen ze diezelfde standaardscore voor een aantal goed van 147 t/m 149.

In 2010 behaalden 3987 leerlingen de hoogste standaardscore. Slechts drie leerlingen hadden alle 200 opgaven van de onderdelen Taal, Rekenen-Wiskunde en Studievaardigheden goed.

De cumulatieve percentages kunnen licht verschillen vanwege afrondingsfouten in de omzetting van de scores. Ze kunnen echter ook verschillen omdat de leerlingen die in 2010 deelnamen aan de Eindtoets in het algemeen net iets vaardiger of iets minder vaardig zijn dan de leerlingen die in 2009 deelnamen. Het eerste lijkt op het eerste gezicht het geval te zijn: in 2009 behaalden de leerlingen een gemiddelde standaardscore van 535.46; in 2010 is het gemiddelde 535.85.

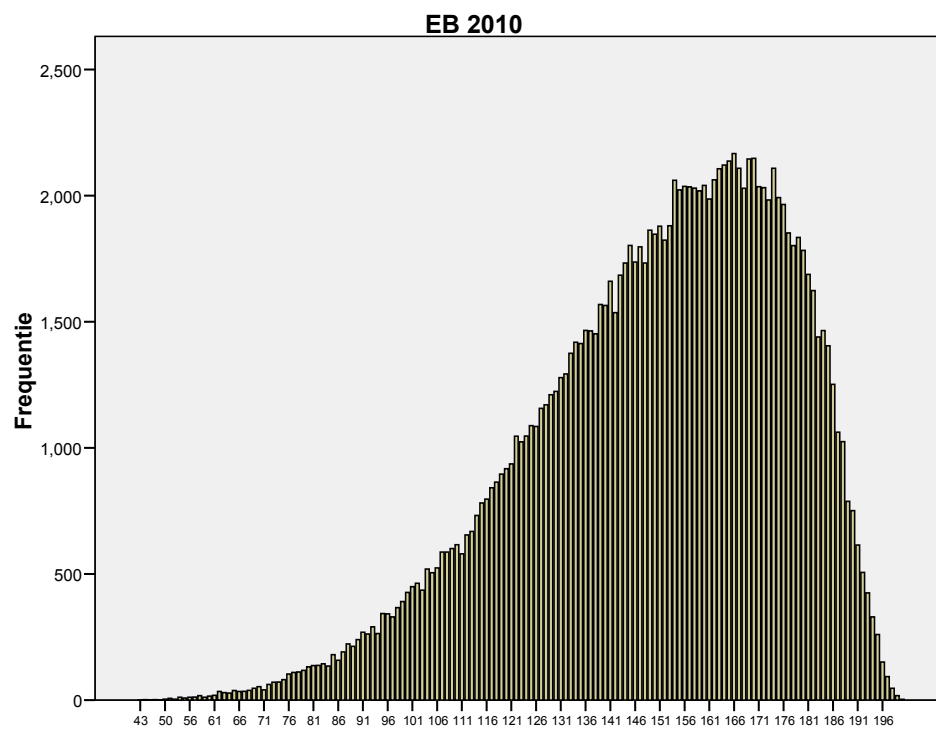
Betekent dit nu dat de leerlingen in 2010 vaardiger zijn dan in 2009? Met deze conclusie dient men voorzichtig te zijn. Er nemen weliswaar veel leerlingen deel aan de Eindtoets, maar dat zijn van jaar tot jaar niet telkens volstrekt vergelijkbare groepen leerlingen. Ook maken meer en meer leerlingen een digitale versie van de Eindtoets. Van de deelnemende scholen in 2009 heeft 6.6% niet meer deelgenomen in 2010. En 4.5% van de in 2010 deelnemende scholen nam in 2009 niet deel aan de Eindtoets. Men kan er niet zonder meer van uitgaan dat de scholen die weggefallen zijn ten opzichte van 2009 vervangen zijn door een 'evenwaardige' groep van scholen in 2010. Omdat het niet zeker is of beide steekproeven in de twee opeenvolgende jaren goed (of even goed) de Nederlandse schoolbevolking vertegenwoordigen, kunnen we ook geen harde uitspraken doen over een daling of stijging van de gemiddelde vaardigheid die door de Eindtoets wordt gemeten.

In figuur 3.1 is de frequentieverdeling van de ruwe totaalscore grafisch weergegeven. Het linksscheve karakter van de verdeling is duidelijk herkenbaar. De vorm van de verdeling voor andere aggregatieniveaus (toetsonderdelen) is in grote lijnen vergelijkbaar.

Tabel 3.3 *Frequenties van scores op de Eindtoets 2009 en de Eindtoets 2010*

Standaard- score	2009			2010		
	van	tot	cum%	van	tot	cum%
501	0	55	0.0	0	54	0.0
502	56	58	0.1	55	57	0.1
503	59	60	0.1	58	60	0.1
504	61	63	0.1	61	62	0.2
505	64	66	0.2	63	65	0.3
506	67	69	0.3	66	68	0.4
507	70	71	0.4	69	71	0.6
508	72	74	0.6	72	73	0.7
509	75	77	0.9	74	76	1.0
510	78	79	1.1	77	79	1.3
511	80	82	1.4	80	82	1.7
512	83	85	1.9	83	85	2.1
513	86	88	2.4	86	87	2.4
514	89	90	2.8	88	90	3.0
515	91	93	3.6	91	93	3.7
516	94	96	4.5	94	96	4.5
517	97	99	5.5	97	99	5.4
518	100	101	6.2	100	101	6.1
519	102	104	7.5	102	104	7.2
520	105	107	8.8	105	107	8.4
521	108	109	9.9	108	110	9.8
522	110	112	11.6	111	113	11.3
523	113	115	13.5	114	115	12.4
524	116	118	15.6	116	118	14.3
525	119	120	17.1	119	121	16.3
526	121	123	19.4	122	124	18.6
527	124	126	22.0	125	126	20.2
528	127	128	23.9	127	129	22.8
529	129	131	26.9	130	132	25.5
530	132	134	30.0	133	135	28.6
531	135	137	33.4	136	138	31.7
532	138	139	35.6	139	140	34.0
533	140	142	39.3	141	143	37.5
534	143	145	43.2	144	146	41.3
535	146	148	47.2	147	149	45.1
536	149	150	49.9	150	152	49.1
537	151	153	54.2	153	154	51.9
538	154	156	58.4	155	157	56.3
539	157	158	61.4	158	160	60.6
540	159	161	65.9	161	163	65.0
541	162	164	70.4	164	165	68.0
542	165	167	74.8	166	168	72.6
543	168	169	77.7	169	171	77.1
544	170	172	81.9	172	174	81.4
545	173	175	85.8	175	177	85.6
546	176	177	88.4	178	179	88.1
547	178	180	91.8	180	182	91.8
548	181	183	94.6	183	185	94.8
549	184	186	96.9	186	188	97.2
550	187	200	100.0	189	200	100.0

Figuur 3.1 Verdeling van de ruwe totaalscore op de Eindtoets 2010



4 Normering

4.1 De Eindtoets: het specifieke karakter van de normering

De functie van de Eindtoets is het geven van onafhankelijke informatie over wat een leerling in acht jaar basisonderwijs heeft geleerd, ten behoeve van de keuze van het brugklastype in een school voor voortgezet onderwijs dat het best bij deze leerling past. Zo bezien vormen in principe alle leerlingen in groep 8 van het basisonderwijs in Nederland de doelgroep van het instrument. Omdat de Eindtoets jaarlijks in een inhoudelijk vernieuwde vorm wordt afgenomen, in dit geval dus in het schooljaar 2009-2010, gaat het om de leerlingen die in dat schooljaar in leerjaar 8 verbleven.

In 2009-2010 namen 137 301 leerlingen deel aan de Eindtoets 2010. Dat is een groot deel van de totale populatie van leerlingen in groep 8 van het reguliere basisonderwijs. Hoeveel dat er exact zijn, is niet te zeggen omdat noch het CBS, noch het CFI gegevens verschaft over aantallen leerlingen per leerjaar. Voor een schatting zou men kunnen uitgaan van het totale aantal leerlingen in een leeftijdscohort, bijvoorbeeld alle elfjarigen. CBS-gegevens (verworven via Statline) leren ons dat het totaal aantal elfjarigen in ons land neerkomt op 200.511. Voor de twaalf- en dertienjarigen zijn dit er respectievelijk 194 826 en 193 568. We kunnen ook kijken naar statistieken over het onderwijs. Volgens het CBS zaten er op de peildatum van 1 oktober 2009 190 030 elfjarigen in het primair onderwijs, van wie 175 768 in het reguliere basisonderwijs. Afhankelijk van het gekozen uitgangsaantal zou men kunnen stellen dat zo'n 70 tot 80% van de leerlingenpopulatie aan de Eindtoets deelneemt. Het heeft weinig zin om exactere aantallen te schatten; de cijfers komen er op neer dat een ruime meerderheid van de leerlingen in de doelgroep meedoet.

Bij het normeren van een test wordt normaliter een representatieve steekproef van voldoende omvang getrokken uit de populatie van leerlingen die nu (of in de toekomst) de doelgroep van de test vormen. Daarbij heeft elk lid van de populatie idealiter dezelfde kans om in die steekproef terecht te komen. Bij de Cito Eindtoets wordt een dergelijke steekproef niet getrokken. Men zou dit kunnen doen, maar het hoeft niet, eenvoudigweg omdat de normen worden vastgesteld nadat de toets bij alle leerlingen is afgenomen en voor alle leerlingen de scores bekend zijn. Bij het omzetten van ruwe scores in standaardscores, dan wel het bepalen van de bijbehorende percentielscores op basis van de cumulatieve frequenties wordt een leerling dus vergeleken met de totale groep leerlingen bij wie de toets in dat schooljaar wordt afgenomen. Iets soortgelijks gebeurt bij het koppelen van de score van een leerling aan het advies over het best passend brugklastype, wat men ook een vorm van normeren zou kunnen noemen. Door middel van het Toelatings- en doorstroomonderzoek (T&DO, zie verder paragraaf 6.3.3) worden de toetsscores van de deelnemende leerlingen gekoppeld aan gegevens (met betrekking tot brugklastype) over de instroom in het eerste leerjaar van het voortgezet onderwijs, in het schooljaar volgend op de toetsafname. Tegenwoordig kan die koppeling voor circa 90% van de deelnemende leerlingen tot stand worden gebracht.

Kortom, de wijze van 'normeren' van de Cito Eindtoets verschilt wezenlijk van de gebruikelijke wijze van het normeren van een test of toets. Daarbij is het vereist dat er een controle wordt uitgevoerd op de representativiteit. Komt de verdeling van de steekproef op een (beperkt) aantal relevante variabelen overeen met de overeenkomstige populatieverdeling? Steekproef en populatie dienen zich op een bepaalde manier tot elkaar te verhouden. In het geval van de Eindtoets is die verhouding op een eigenaardige wijze gedefinieerd, afhankelijk van het perspectief dat men inneemt:

- Men gaat ervan uit dat de normering plaatsvindt op basis van de volledige populatie van Eindtoetsgebruikers. De score van een leerling wordt dus gerelateerd aan de scoreverdeling binnen deze gehele populatie. Er is in dit perspectief geen steekproef. Wel moet men zich afvragen hoe deze populatie van gebruikers zich verhoudt tot de totale doelgroepopulatie.
- Men beschouwt de groep van Eindtoetsgebruikers als een steekproef uit de totale doelgroep (doelpopulatie). In dit perspectief is er wél een steekproef, maar is het zeker dat dit geen aselechte steekproef is. Voor alle leerlingen uit de doelgroep geldt immers dat zij niet per definitie dezelfde kans

hebben om in deze steekproef terecht te komen. Alleen leerlingen die op scholen zitten die deelnemen aan de Eindtoets hebben een kans, waarbij de scholen zichzelf hebben geselecteerd als gebruiker. De vraag is dan of in deze zelfselectie factoren een rol spelen die ervoor zorgen dat de groep van Eindtoetsgebruikers op een systematisch andere wijze is samengesteld als de totale doelgroepopulatie.

Welk perspectief men ook kiest, het is interessant en belangrijk om inzicht te hebben in de verdelingskenmerken van de groep Cito Eindtoetsgebruikers en die van de totale doelgroep. Men kan dit nagaan op het niveau van de scholen en op het niveau van de leerlingen zelf. In de volgende paragraaf zullen we daarom proberen antwoord te geven op de vraag of de deelpopulatie (dan wel steekproef) van leerlingen bij wie de Eindtoets is afgenomen op een aantal belangrijke kenmerken op dezelfde wijze is verdeeld als de gehele doelgroepopulatie (van leerlingen in leerjaar 8 van het reguliere basisonderwijs in 2010).

4.2 Kenmerken van Eindtoetsgebruikers in vergelijking met de gehele doelpopulatie

In deze paragraaf besteden we aandacht aan de kenmerken die normaliter relevant worden geacht om de relatie tussen normeringssteekproef en populatie in kaart te brengen. Het COTAN-beoordelingssysteem (Evers et al., 2010) geeft in dit opzicht aan dat de normeringssteekproef in ieder geval moet worden beschreven in termen van leeftijd, sekse, etniciteit en regio. Van de variabelen die daaraan volgens het beoordelingssysteem zouden kunnen worden toegevoegd, lijken urbanisatiegraad, sociaaleconomische status en opleidingsniveau (in dit geval van de ouders) het meest in aanmerking te komen.

De verdeling naar *regio* gaat uit van vier regioniveaus die overeenkomen met de volgende indeling naar provincies (CBS-indeling naar landsdeel):

- Noord: Groningen, Friesland en Drenthe;
- Oost: Gelderland, Overijssel en Flevoland;
- West: Utrecht, Noord- en Zuid-Holland, Zeeland;
- Zuid: Noord-Brabant en Limburg.

De verdeling naar urbanisatiegraad of *mate van verstedelijking* is terug te voeren op de bij het CBS gebruikelijke indeling naar vijf niveaus, namelijk zeer sterk, sterk, matig, weinig en niet verstedelijkt. Verder kan er in dit verband ook gekeken worden naar schoolgrootte. Daarbij wordt onderscheid gemaakt tussen grote scholen (≥ 200 leerlingen) en kleine scholen (< 200 leerlingen).

De verdeling naar *seks* levert geen probleem op.

Wat betreft *leeftijd* gaat het hier natuurlijk om een betrekkelijk homogene groep 11- à 12-jarige leerlingen die op een bepaald moment in groep 8 collectief de Eindtoets maakt. Daarbij is er sprake van de gebruikelijke spreiding van leeftijden over een schooljaar, met dien verstande dat sommige leerlingen voorlijk genoemd kunnen worden (bijvoorbeeld omdat ze ooit een klas hebben overgeslagen) en andere vertraagd (omdat ze bijvoorbeeld een of meerdere malen hebben gedoubleerd). Deze verdeling naar leeftijd zal worden beschreven zonder dat vergelijking met de totale doelgroepopulatie mogelijk is (omdat gegevens daarover niet beschikbaar zijn).

Wat betreft *etniciteit* is het voor de vergelijking een probleem dat geen eenduidige vergelijkingsgegevens voor de populatie bekend zijn. Voorheen kon er gewerkt worden met een bij Cito gebruikelijke indeling van scholen in drie groepen of strata op basis van een weging van de formatiegewichten van hun leerlingen op schoolniveau. Het formatiegewicht van de leerlingen werd gebruikt voor de bepaling van de formatieomvang van een school. Daarbij werden de leerlingen gecategoriseerd in vijf formatiegewichten die een combinatie vormden van opleidingsniveau, sociaaleconomische status en etnische herkomst van de ouders. Inmiddels is echter de categorisering naar leerlinggewichten vervangen door een andere regeling waarbij nog slechts drie niveaus worden onderscheiden en waarbij etniciteit geen rol meer speelt. De definities van de categorieën zijn aangescherpt en geheel gebaseerd op het opleidingsniveau van de ouders.

Gewicht	Omschrijving
0	één van de ouders of beide ouders heeft of hebben een opleiding gehad uit categorie 3 (zie voor de categorieomschrijving hieronder)
0.3	beide ouders of de ouder die belast is met de dagelijkse verzorging heeft of hebben een opleiding uit categorie 2 gehad
1.2	één van de ouders heeft een opleiding gehad uit categorie 1 en de ander een opleiding uit categorie 1 óf 2

Categorie	Omschrijving
1	maximaal basisonderwijs of (v)so-zmlk
2	maximaal lbo/vbo, praktijkonderwijs of vmbo basis- of kaderberoepsgerichte leerweg
3	overig vo en hoger

Dit betekent dat we nog wel gebruik kunnen maken van gegevens over de *achterstandssituatie* van leerlingen in termen van het opleidingsniveau van de ouders (als indicator voor sociaaleconomische status), maar dat een verwijzing naar ethnische herkomst ontbreekt.

Wel is navraag gedaan naar de taal die door de leerling thuis wordt gesproken (de *thuis taal*). Voor zover etniciteit van belang is voor leervorderingen, lijken deze gegevens over thuis taal een goed alternatief. Vergelijking met de doelgroep populatie is echter niet mogelijk omdat de betreffende gegevens ontbreken. Er worden drie niveaus onderscheiden, namelijk (thuis) Nederlandssprekend, anderstalig en een combinatie van Nederlands met een andere taal.

In het navolgende worden regio, mate van verstedelijking en schoolgrootte op schoolniveau besproken. Daarnaast worden de scholen gekarakteriseerd en vergeleken in termen van het percentage achterstandsleerlingen (dat wil zeggen, het percentage leerlingen met een afwijkend leerlinggewicht). De kenmerken regio, sekse, leeftijd (leertijd), achterstandssituatie (leerlinggewicht) en thuis taal worden op leerlingniveau aan de orde gesteld.

Schoolkenmerken

In tabel 4.1 is de verdeling gegeven van de Nederlandse populatie van (7066) basisscholen. De gegevens zijn gebaseerd op cijfers van het CFI (het informatie-orgaan van het ministerie van OCW) en op de postcoderelatietafel van Cendris. Voor 43 scholen ontbraken CFI-gegevens en/of gegevens over provincie en mate van verstedelijking. Deze scholen zijn verwijderd zodat een bestand van 7013 scholen overbleef. Van deze 7013 scholen namen er 5674 (80.9%) deel aan (de papieren versie van) de Eindtoets Basisonderwijs 2010. Alle gegevens hebben betrekking op het schooljaar 2009-2010.

In de tabel zijn de verdelingen gegeven naar regio, schoolgrootte, mate van verstedelijking en percentage achterstandsleerlingen, voor zowel de populatie van basisscholen als voor de subgroep van scholen die aan de Eindtoets deelnemen.

Tabel 4.1 Verdeling van basisscholen naar regio, mate van verstedelijking, schoolgrootte en percentage achterstandsl leerlingen voor de Nederlandse populatie en voor de subgroep van Eindtoetsgebruikers.

	Regio			Schoolgrootte	
	Populatie (%)	EB 2010 (%)		Populatie (%)	EB 2010 (%)
noord	15.9	15.2	klein	49.5	49.2
oost	24.2	24.7	groot	50.5	50.8
west	41.1	39.9			
zuid	18.8	20.3	totaal	100	100
totaal	100	100			

	Mate van verstedelijking			Percentage achterstandsl leerlingen	
	Populatie (%)	EB 2010 (%)		Populatie (%)	EB 2010 (%)
zeer sterk	11.7	12.6	0 – 10%	49.8	48.1
sterk	20.9	19.6	11 – 25%	31.1	31.7
matig	20.3	19.9	26 – 50%	12.0	12.3
weinig	26.7	27.7	51 – 100%	7.1	7.9
geen	20.4	20.2			
			totaal	100	100
totaal	100	100			

Het is duidelijk dat de subgroep van scholen die de Eindtoets 2010 hebben afgenomen bij hun leerlingen een nauwkeurige afspiegeling vormt van de totale populatie van basisscholen in Nederland. Dat geldt voor alle besproken achtergrondkenmerken.

Leerlingenkenmerken

Om de verdeling van de deelnemersgroep aan de Eindtoets 2010 voor de variabelen regio, sekse, leertijd (leeftijd), thuistaal en leerlinggewicht te kunnen evalueren werd de populatieverdeling zo goed mogelijk beschreven of geschat op basis van beschikbare gegevens. De bron daarvan was niet in alle gevallen dezelfde omdat dergelijke gegevens niet altijd systematisch voor alle variabelen beschikbaar zijn.

Gegevens werden ontleend aan CBS en CFI en betroffen soms de populatie leerlingen van 10 tot 14 jaar, soms de subgroep van elfjarigen, dit afhankelijk van de beschikbaarheid van gegevens. Voor de schatting van de populatieverdeling werden steeds de referentiegegevens gebruikt die het meest van toepassing waren. Voor sommige variabelen is de verdeling in de populatie niet bekend

Regio

De deelnemers aan de Eindtoets 2010 waren voor 9.5% afkomstig uit het landsdeel Noord (10.4%), 23.8% kwam uit landsdeel Oost (22.8%), 43.7 % uit landsdeel West (45.7%) en 23.7% uit landsdeel Zuid (21.2%). Tussen haakjes de percentages elfjarigen onderverdeeld naar landsdeel (CBS Statline voor 2010).

Gegeven het feit dat de subpopulatie van elfjarigen niet volledig samenvalt met de groep leerlingen die leerjaar 8 van het reguliere basisonderwijs bevolkt, kan worden geconstateerd dat de regionale verdeling van Eindtoetsgebruikers nagenoeg gelijk is aan de betreffende verdeling van de totale populatie.

Sekse

Volgens CBS-tellingen over 2009 (CBS Statline) bestaat 51.2% van de populatie 10- tot 14-jarigen uit jongens en 48.8% uit meisjes. Deze leeftijdsgroep is wel iets ruimer dan de leeftijdsgroep die groep 8 van het reguliere basisonderwijs bevolkt, maar omdat de CBS-gegevens over de afgelopen jaren uiterst stabiel waren, mag worden aangenomen dat deze verdeling ook voor leerlingen van groep 8 geldt. Eveneens volgens het CBS zijn de aantallen jongens en meisjes in de leeftijdsgroep elfjarigen in het reguliere basisonderwijs nagenoeg gelijk verdeeld (Statline; jongens 49.97%, meisjes 50.03%). De lichte

oververtegenwoordiging van meisjes is grotendeels te wijten aan het feit dat meer jongens dan meisjes op scholen voor speciaal onderwijs en speciaal basisonderwijs verblijven. Voor de groep deelnemers aan de Eindtoets (die niet samenvalt met de groep van elfjarigen; scholen voor speciaal (basis)onderwijs behoren meestal niet tot de deelnemers) treffen we een soortgelijke verdeling aan: 49.7% jongens en 50.3% meisjes.

Leeftijd/leertijd

Op basis van de leeftijd van de leerlingen is nagegaan welke leerlingen op de peildatum van 1 oktober 2009 12 jaar of ouder waren. Van deze leerlingen mag worden aangenomen dat zij langer dan op basis van hun leeftijd noodzakelijk was in de kleutergroep zijn gehouden en/of dat zij in het verdere verloop van hun schoolloopbaan zijn blijven zitten. De gegevens met betrekking tot de EB-afname in 2010 laten zien dat ongeveer 18.4% van de leerlingen vertraging heeft opgelopen (20.5% van de jongens en 16.3% van de meisjes; zie voor een bespreking van dit verschil hoofdstuk 6 onder begripsvaliditeit). Voor dit kenmerk zijn geen populatiegegevens beschikbaar. Wel kan worden vastgesteld dat deze cijfers vergelijkbaar zijn met de resultaten uit het periodieke peilingsonderzoek (PPON) dat met regelmaat door Cito wordt uitgevoerd (zie bijvoorbeeld Heesters, Van Berkel, Van der Schoot & Hemker, 2007). Dit type peilingsonderzoek, dat landelijk representatief is, laat een stabiel beeld zien voor groep 8, namelijk dat ongeveer 20% van de leerlingen als vertraagd kan worden aangemerkt volgens het hierboven beschreven criterium.

Thuis taal

Ten aanzien van de door leerlingen gesproken thuistaal wordt onderscheid gemaakt tussen leerlingen die thuis uitsluitend Nederlands spreken, leerlingen die thuis daarnaast nog een andere taal spreken (Nederlands+) en leerlingen die thuis uitsluitend een andere taal spreken en geen Nederlands. Van de leerlingen die aan de Eindtoets 2010 deelnamen, sprak 89.1% thuis (uitsluitend) Nederlands. Daarnaast sprak 4% van de deelnemers thuis naast het Nederlands nog een andere taal. De overige 6.9% sprak thuis uitsluitend in een andere taal, van wie 2.2% Turks, 1.7% Arabisch, 0.4% Surinaams of Antilliaans en 0.6% een andere West-Europese taal dan Nederlands; nog eens 1.8% sprak thuis een andere taal dan de al genoemde. Voor de totale doelgroep van leerlingen in het achtste leerjaar van het reguliere basisonderwijs zijn soortgelijke gegevens niet bekend.

Leerlinggewicht

Voor de deelnemers aan de Eindtoets heeft de school in bijna 84% van de gevallen aangegeven wat hun leerlinggewicht is. Van deze ruim 115 000 leerlingen had 84.1% (84.6%) geen afwijkend leerlinggewicht, of, met andere woorden, een leerlinggewicht van 0. 8.9% (6.5%) had een leerlinggewicht van 0.3 en nog eens 6.9% een leerlinggewicht van 1.2 (8.9%). Deze percentages zijn vergelijkbaar met populatiegegevens die het CFI verstrekt (deze staan weer tussen haakjes vermeld), zeker als men het grote aantal ontbrekende waarden in de dataset in aanmerking neemt, evenals het feit dat de CFI-data betrekking hebben op alle leerlingen die het basisonderwijs bevolken (dus niet alleen leerlingen in leerjaar 8).

Conclusie

Op basis van de hierboven gepresenteerde gegevens is de conclusie gerechtvaardigd dat de totale groep leerlingen die heeft deelgenomen aan de Cito Eindtoets Basisonderwijs zeer sterk lijkt op de totale populatie leerlingen die leerjaar 8 van het reguliere basisonderwijs bevolkt, althans voor wat betreft een aantal belangrijke kenmerken zoals sekse, regio, mate van verstedelijking en leerlinggewicht. Voor de deelnemende scholen kan hetzelfde worden geconstateerd in vergelijking met de landelijke populatie van scholen voor regulier basisonderwijs. De kans is daarmee zeer gering dat zelfselectie de populatie van Eindtoetsgebruikers heeft gemaakt tot een specifieke subpopulatie die op relevante factoren afwijkt. Het feit dat de deelpopulatie van Eindtoetsgebruikers zo'n grote omvang heeft in verhouding tot de gehele populatie draagt sterk aan deze conclusie bij.

Anders gezegd: de groep Eindtoetsgebruikers kan desgewenst worden beschouwd als een (zeer grote) representatieve steekproef uit de totale populatie van leerlingen in groep 8 van het reguliere basisonderwijs. Bij uitspraken als 'bij een ruwe score van 178 goed op de Eindtoets van leerling Ingrid hoort een standardscore van 546 en als best daarbij passend advies brugklastype vwo' kan men dus de totale Nederlandse populatie aanhouden als referentiekader.

4.3 Kalibratie en normering

In hoofdstuk 2 zijn in algemene zin de procedures beschreven die in eerste instantie leiden tot gekalibreerde opgavenbanken en in tweede instantie tot versies van de Eindtoets die over verschillende jaargangen gezien volledig equivalent zijn. Tevens is daar ingegaan op het meetmodel dat ten grondslag ligt aan de Eindtoets. In deze paragraaf gaan we nog wat nauwkeuriger in op de stappen die in het kader van de kalibratie zijn genomen (4.3.1) en geven we resultaten van 1 kenmerken van de verdeling aan de orde geweest. In paragraaf 4.3.3 kunnen we ons daarom beperken tot de beschrijving van adviescategorieën die zijn vastgesteld voor verschillende uitkomsten van de Eindtoets en wel op basis van Toelatings- en doorstroomonderzoek.

4.3.1 De stappen in de kalibratie

Met kalibratie wordt bedoeld dat we kengetallen zoeken bij de items die de antwoorden van de leerlingen goed representeren. Hoe de kengetallen gezocht worden ligt deels vast door het gekozen model (zie paragraaf 2.4.2.2) en hoe succesvol deze operatie is, kan statistisch getoetst worden. Eenvoudig gezegd, schatten we in OPLM met de CML-methode de itemparameters en controleren we of deze de data goed voorspellen. Voor een exacte beschrijving van de statistische toetsen die in OPLM gebruikt worden, hun eigenschappen en feitelijke implementatie in OPLM verwijzen we naar Verhelst (1993). Hier beperken we ons tot een korte beschrijving van de principes van de statistische toetsen die gebruikt zijn in de kalibratieprocedure. De statistische toetsen in OPLM hebben goede statistische en asymptotische eigenschappen daar OPLM behoort tot de exponentiële familie, met de gewogen somscore,

$$s = \sum_{i=1}^k a_i x_i, \quad (4.1)$$

als een 'afdoende statistiek' voor de vaardigheid θ . Dit betekent dat alle informatie in de data met betrekking tot de vaardigheid in deze statistiek aanwezig is. Hiervan wordt gebruikgemaakt bij de statistische toetsen in OPLM. Het basisprincipe van de statistische toetsen in OPLM is dat op grond van de afdoende statistiek s de personen in de data kunnen worden gegroepeerd.

En binnen deze groepen kan de verwachte proportie goede antwoorden op een item onder het model, $p(+|s)$, vergeleken worden met de feitelijk geobserveerde proportie goede antwoorden, $prop(+|s)$. In het polytome geval worden de items gedichotomiseerd, de proportie goede antwoorden verwijst dan naar de hoge itemscore (zie Verhelst, 1993, hoofdstuk 7). Via de basisvergelijking van OPLM kunnen we eenvoudig de conditionele kans op het goed beantwoorden van de items afleiden en daarmee kunnen we $p(+|s)$ evalueren, $prop(+|s)$ volgt uit de data. Discrepancies tussen $p(+|s)$ en $prop(+|s)$ duiden op schendingen van het model. Deze discrepanties vormen de basis voor de diverse statistische toetsen in OPLM. De toetsingsgrootte voor de veronderstelde discriminatie-indices is gegeven door

$$M = f_{s \in H}(p(+|s) - prop(+|s)) + f_{s \in L}(prop(+|s) - p(+|s)). \quad (4.2)$$

Deze zogenaamde M-toetsen verdelen de scoregroepen in een laag deel (L) en een hoog deel (H) en f is een monotone functie. M-toetsen hebben een duidelijke interpretatie: is M significant positief dan is de veronderstelde steilheid van de ICC (item karakteristieke curve) overschat in het model, is M daarentegen erg laag dan is de index te klein. Verhelst laat zien voor welke functie, f , $M \approx N(0,1)$. In OPLM zijn drie verschillende M-toetsen geïmplementeerd die verschillen in de definitie van de hoge en lage scoregroepen. Naast deze M-toetsen is er een algemene itemtoets die de volgende vorm heeft

$$S = f(p(+|s) - prop(+|s)).$$

Deze zogenaamde S-toets heeft een χ^2 verdeling onder het model. Analoog hieraan is er ook een toets om vormen van vraagonzuiverheid (in het Engels 'item bias' of 'differential item functioning', afgekort DIF) op te sporen:

$$S = h(p_I(+|s) - prop_I(+|s), (p_{II}(+|s) - prop_{II}(+|s))),$$

waarbij *I* en *II* de twee niveaus van de variabele indiceren waarvoor we de bias onderzoeken.

Als globale modeltoets is de R1c-toets (Glas, 1988) geschikt. Ook de distributie van alle afzonderlijke S-toetsen komt hiervoor in aanmerking. Als we deze S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn op het (0,1) interval. Kortom, als we afzien van de formeel-statistische achtergrond van de gehanteerde toetsen, kan de kalibratieprocedure als volgt worden samengevat:

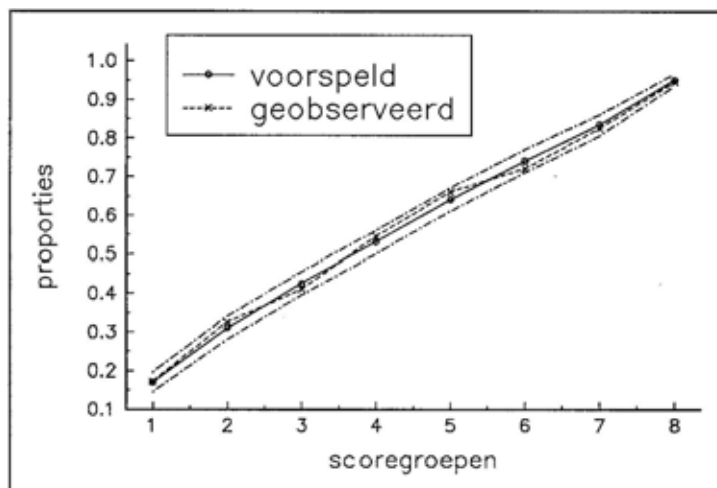
- 1 Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
- 2 Vervolgens schatten we de itemparameters met behulp van de CML-methode.
- 3 Met behulp van de M-toetsen controleren we of de discriminatie-indices goed zijn ingesteld.
- 4 Een volgende controle betreft de overschrijdingskansen van de S-toetsen en een grafische modelcontrole door middel van het programma OPDRAW (grafische inspectie van de ICC's).
- 5 Vervolgens vindt een globale modelcontrole plaats in de vorm van een R1c-toets en de verdeling van de overschrijdingskansen van de S-toetsen.
- 6 Daarna toetsen we nadrukkelijk op vormen van item bias, in dit geval naar geslacht en herkomst (thuisstaat) en leerlinggewicht.

De stappen 1 tot en met 6 worden een aantal malen doorlopen tot het resultaat bevredigend is. Afhankelijk van de uitkomsten kunnen items worden verwijderd. Ook inhoudelijke overwegingen (zie hiervoor hoofdstuk 2 over de achtergronden van de toetsinhoud) spelen een rol in dit beslissingsproces. De opgaven vormen na de kalibratie een gekalibreerde opgavenbank, waarbij de opgaven per onderscheiden vaardigheidsdimensie een beroep doen op hetzelfde complex aan vaardigheden of 'latente trek'.

4.3.2 Evaluatie van de kalibratieprocedure

Het is niet eenvoudig om de kwaliteit van de kalibratie aan te tonen. De belangrijkste statistische instrumenten om de passing van een opgave in het IRT-model te bewerkstelligen en uiteindelijk te documenteren betreffen de hierboven al besproken S-toetsen. Het lastige daarvan is, dat de toetsing voor een groot deel visueel gebeurt. Dit kunnen we illustreren aan de hand van figuur 4.1 (zie Staphorsius, 1994, blz. 239). Figuur 4.1 beeldt voor een opgave de gegevens af waarop de betreffende S -toetsen gebaseerd zijn (zie handleiding OPLM: Verhelst, 1992). Ten behoeve van deze toetsing wordt de totale groep van leerlingen die een verzameling opgaven gemaakt heeft, ingedeeld in een aantal (meestal acht) scoregroepen. Elke groep bestaat uit leerlingen met een ongeveer even hoge score. De geobserveerde proporties juiste antwoorden van deze groepen (telkens gesymboliseerd door een x) zijn door de middelste stippellijn verbonden. De volle lijn daarentegen verbindt de proporties die op grond van de parameterschattingen voorspeld kunnen worden. De twee buitenste lijnen geven het 95%-betrouwbaarheidsinterval aan. De breedte van dit interval is in belangrijke mate afhankelijk van het aantal leerlingen dat de opgave heeft beantwoord. Uit de figuur blijkt heel duidelijk dat de geobserveerde proporties, zoals bedoeld, binnen het 95%-betrouwbaarheidsinterval van de (geschatte) voorspelde proporties liggen, en dit komt in grote lijnen overeen met een niet-significante S-toetsingsgrootheid (Verhelst, et al., 1994).

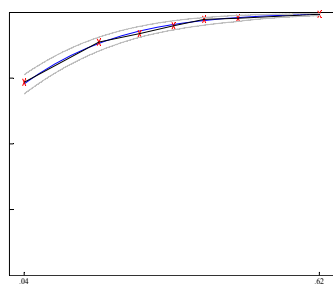
Figuur 4.1 Grafische voorstelling van een S-toets



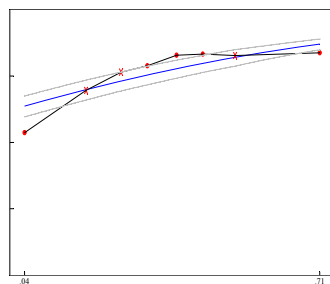
Het is ondoenlijk om voor alle 200 opgaven (290 inclusief Wereldoriëntatie) dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons steeds per toetsonderdeel tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.2 illustreren dat zelfs bij de slechtst passende opgaven sprake is van een zeer aanvaardbaar beeld. Er wordt in deze gevallen voor een deel (van de onderscheiden scoregroepen) niet beantwoord aan de eis dat de geobserveerde proportie binnen het 95%-betrouwbaarheidsinterval van de geschatte proporties ligt. Dit beeld doet zich per toetsonderdeel slechts bij enkele opgaven voor die dan ook een uitzondering vormen. De overige opgaven voldoen voor alle scoregroepen wel aan die eis. De afbeeldingen voor de representatieve en best passende opgaven illustreren dit. Dit leidt tot de conclusie dat bij vrijwel alle opgaven in de Eindtoets een grafische voorstelling van de S-toetsing hoort die in grote lijnen met figuur 4.1 overeenkomt; andere opgaven zijn bij de kalibratie niet in de itembank opgenomen. Dit is, zeker gezien de relatief grote aantallen observaties die in het geding zijn, een zeer sterke aanwijzing dat het meetinstrument en het meetmodel dat ontwikkeld respectievelijk gebruikt is, adequaat zijn om het gedrag van de leerlingen te verklaren. Bovendien blijkt, en dat is vanuit theoretisch oogpunt nog belangrijker, dat gemeten verschillen in gedrag tussen de leerlingen te verklaren zijn door één unidimensionaal concept.

Figuur 4.2 Voorbeelden van S-toetsen voor de Eindtoets Basisonderwijs 2010 met per onderdeel de best passende, de slechtst passende en een qua passing representatieve opgave

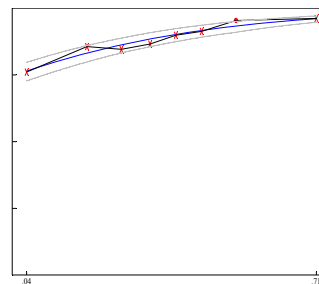
Taal



Best passend

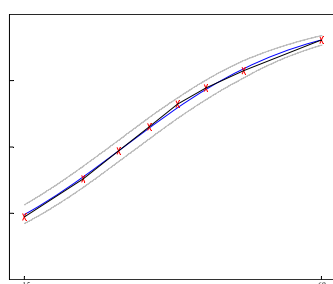


Slechtst passend

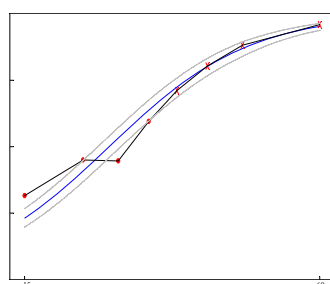


Representatieve passing

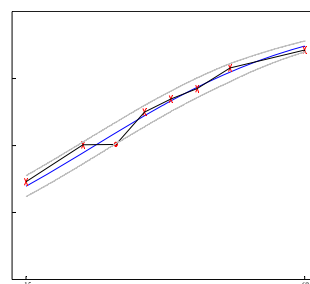
Rekenen-Wiskunde



Best passend

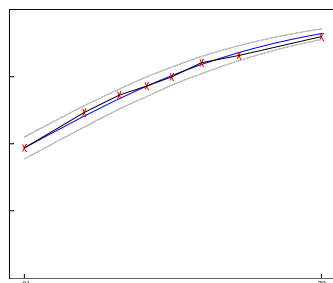


Slechtst passend

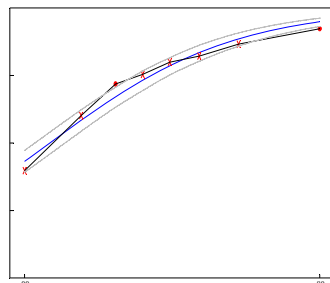


Representatieve passing

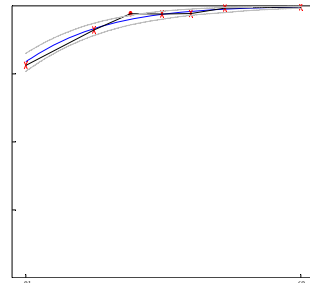
Studievaardigheden



Best passend

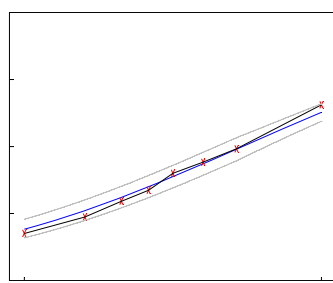


Slechtst passend

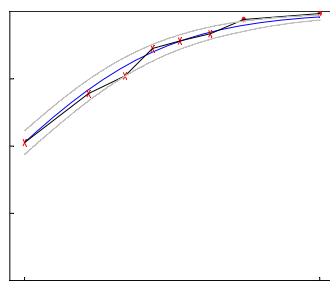


Representatieve passing

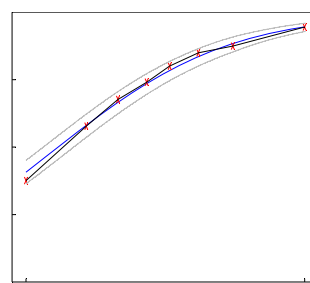
Wereldoriëntatie



Best passend



Slechtst passend



Representatieve passing

In feite kan men bij de kalibratie beter varen op deze grafische weergaven dan op toetsingsresultaten in termen van exacte getallen en de significantie daarvan. Vanwege de zeer grote aantallen (voor de kalibratie in 2010 werd een representatieve aselechte steekproef gebruikt van N=5123 getoetste leerlingen) leidt immers zelfs de minste afwijking van de gewenste situatie (zelfs als de lijn die de geobserveerde proporties representeert geheel binnen het 95%-betrouwbaarheidsinterval ligt) tot significante toetsingsresultaten. Het heeft dus weinig zin om daaraan conclusies te verbinden.

Om toch zinvolle resultaten op de S-toetsen te kunnen laten zien, is een aantal representatieve steekproeven van steeds N=512 (i.e. 10 % van de omvang van de kalibratiesteekproef) uit de data getrokken; bij deze steekproefomvang is de power van de statistische toetsingen optimaal. Voor deze steekproeven zijn de S-toetsen opnieuw uitgevoerd waarbij we vooral geïnteresseerd zijn in de distributie van de overschrijdingskansen van deze verzameling toetsingsresultaten. Als we de S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn binnen het (0,1) interval, uiteraard met zo weinig mogelijk significante resultaten. Tabel 4.2, waarin het (0,1) interval is opgedeeld in tien gelijke stukken, geeft een beeld van de uitkomsten voor tien verschillende steekproeven bij een kalibratie alsof alle opgaven van de Eindtoets op één vaardigheidsdimensie zouden liggen. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan .01, respectievelijk .05. Het is duidelijk dat de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Dat is zelfs zo bij de steekproef die de slechtst passende verdeling laat zien, de op één na laatste steekproef. Deze resultaten geven een bevestiging van het eerder geschetste beeld dat, met uitzondering van enkele opgaven, sprake is van niet-significante S-toetsen. Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensionaal construct representeren.

Tabel 4.2 Verdeling van overschrijdingskansen bij S-toetsen bij 10 aselechte steekproeven van N=512 uit de kalibratiesteekproef bij alle opgaven van de EB 2010 (exclusief Wereldoriëntatie)

0.-/---/-----.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1.											
3	8	12	14	18	22	15	15	19	20	21	32
1	7	6	18	15	27	20	21	18	24	17	25
0	3	8	26	17	13	28	22	21	15	25	21
2	3	8	14	26	16	17	19	20	23	20	30
3	9	5	17	13	13	26	24	19	18	30	22
1	6	9	23	18	17	17	28	27	11	21	22
0	6	9	23	18	17	17	28	27	11	21	22
0	9	8	13	27	17	24	18	24	19	13	27
3	4	12	14	16	18	28	18	24	25	20	17
2	4	11	16	26	17	14	23	21	22	20	23

NB. Het betreft 199 opgaven, bij één opgave is de moeilijkheidsgraad zodanig dat een S-toets niet mogelijk was.

Wat de slechtst passende verdeling is, is te zien in tabel 4.3. In deze tabel zijn de R1c-waarden weergegeven voor dezelfde tien steekproeven waarvoor in tabel 4.2 de resultaten van de S-toetsen zijn weergegeven. R1c is een statistiek die zicht geeft op de modelpassing van de toets (of een toetsonderdeel) als geheel. Ook voor deze statistiek is het nodig om ons te baseren op zo'n kleinere steekproef (met adequate statistische power) om zinnige resultaten te verkrijgen. Voor een acceptabele modelfit geldt als vuistregel dat R1c bij voorkeur niet significant zou moeten zijn en niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden.

Het is duidelijk dat de modelpassing van de toets vrijwel optimaal is. In alle tien steekproeven is R1c kleiner dan het aantal vrijheidsgraden, terwijl de p-waarde nergens significant is. Sterker nog, bij sommige steekproeven is deze p-waarde zelfs gelijk aan 1. Daarbij dient men te bedenken dat ter berekening van de R1c een virtuele kalibratie is uitgevoerd waarbij is gesimuleerd dat alle opgaven één vaardigheidsdimensie representeren. De eigenlijke kalibratie is uitgevoerd op één of meer vaardigheidsdimensies per toetsonderdeel. Zo is er bij het onderdeel Rekenen-Wiskunde sprake van één vaardigheidsdimensie, terwijl

bij Taal onderscheid wordt gemaakt naar afzonderlijke vaardigheidsdimensies voor Begrijpend lezen, Schrijven, Woordenschat en Spelling. Dit houdt in dat de modelpassing per vaardigheidsdimensie in feite nog optimaler is dan de resultaten in de tabellen 4.2 en 4.3 al suggereren.

Ten slotte bespreken we nog een methode om de modelpassing te verantwoorden die wordt besproken in het COTAN Beoordelingssysteem (COTAN 2010, p. 40). Het betreft hier een poging om de nauwkeurigheid van de itemparameterschattingen te beoordelen op basis van een constante (in het COTAN-Beoordelingssysteem met 'c' aangeduid) die weergeeft hoe de relatie is tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie. Het beoordelingssysteem geeft ook richtlijnen voor het beoordelen van de grootte van deze 'c'. Deze dient te worden beoordeeld als goed als de waarde lager is dan of gelijk aan .20. Waarden tussen .30 en .40 kunnen nog als voldoende worden beschouwd.

Tabel 4.3 R1c-waarden bij 10 aselecte steekproeven van N=512 uit de kalibratiesteekproef bij alle opgaven van de EB 2010 (exclusief Wereldoriëntatie)

Steekproef #	R1c	df	p
1	531.068	597	.974
2	567.264	597	.804
3	580.370	597	.681
4	542.586	597	.945
5	518.957	597	.990
6	462.135	597	1.000
7	473.220	597	1.000
8	561.137	597	.851
9	580.777	597	.677
10	555.288	597	.888

De waarden voor deze constante zijn weergegeven in tabel 4.4, zowel voor de toets als geheel als voor de onderdelen (inclusief Wereldoriëntatie). De gemiddelde waarden van de constante zijn uitstekend te noemen. Ze zijn alle lager dan .20 en variëren tussen .033 en .051. Voor geen enkele opgave is c groter dan .20. De conclusie mag luiden dat we ook op basis van deze analyse de kalibratie geslaagd kunnen noemen.

Tabel 4.4 Nauwkeurigheid van de itemparameterschattingen (constante 'c') voor de Eindtoets als geheel en voor de toetsonderdelen afzonderlijk (inclusief Wereldoriëntatie)

Toets(onderdeel)	Constante 'c'	
	Range	Gemiddelde
Eindtoets 2010 totaalscore	.024 - .137	.047
Taal	.024 - .095	.048
Rekenen-Wiskunde	.018 - .066	.033
Studievaardigheden	.027 - .127	.048
Wereldoriëntatie	.030 - .186	.051

Wat betreft de laatste stap in de kalibratieprocedure: er zijn DIF-analyses uitgevoerd voor alle opgaven van de Eindtoets en wel voor sekse en thuistaal. De uitkomsten van deze analyses worden afzonderlijk besproken in hoofdstuk 6 onder de noemer begripsvaliditeit.

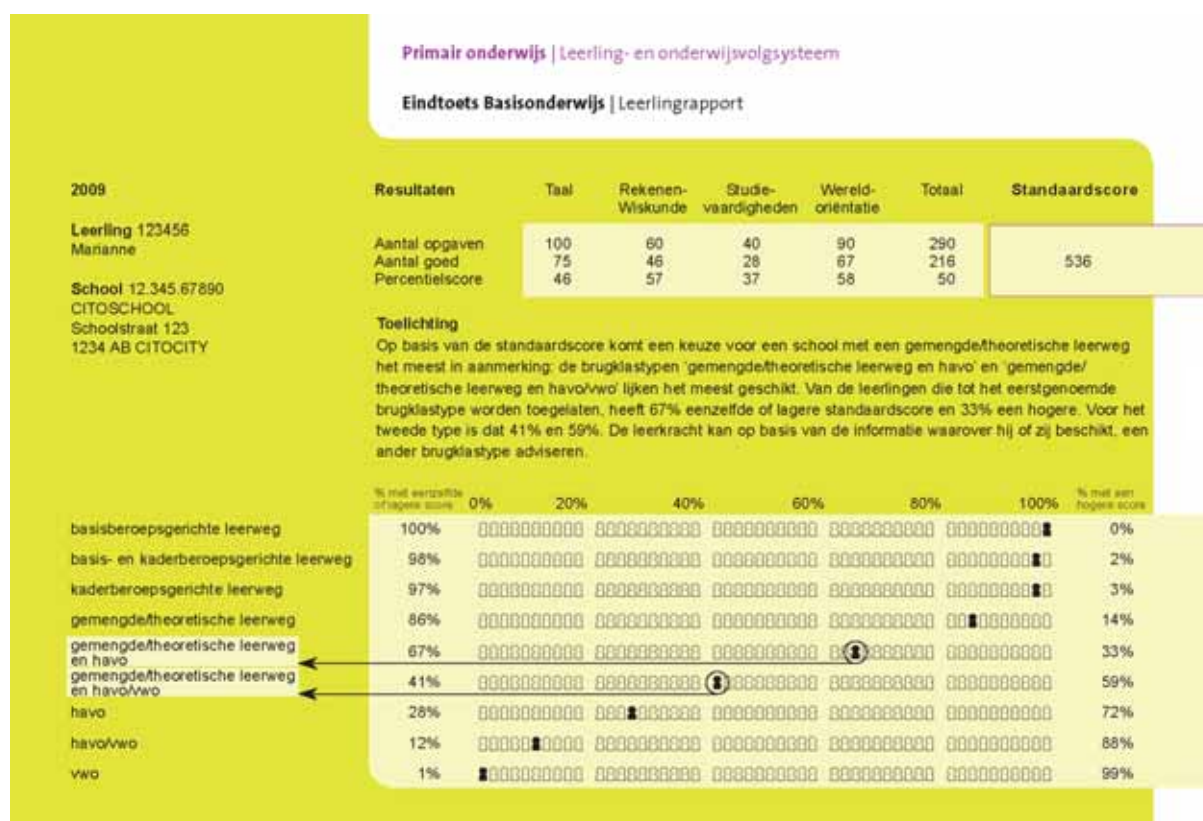
Conclusie

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de Eindtoets en de onderdelen daarvan de kalibratie geslaagd is. Hiermee is het laatste woord nog niet gezegd over de validiteit, maar het kalibratieonderzoek brengt in ieder geval een essentieel aspect van het validiteitsvraagstuk naar voren: de rechtvaardiging van wat in de meeste toetstoepassingen gebruikelijk is, namelijk het reduceren van alles wat de leerling heeft geantwoord tot een enkele toetsscore (of afgeleid daarvan, een enkele schatting van zijn onderliggende vaardigheid). De kalibratie-analyse, als puur formeel proces, kan geen uitspraken doen over de inhoudsvaliditeit of over de constructvaliditeit als antwoord op de vraag: hoe kan worden aangetoond dat het concept dat de items in de bank meten dekkend is voor en samenvalt met het construct dat we in de Eindtoets proberen te meten (zoals dat in het didactisch en het wetenschappelijk forum wordt bedoeld)? In hoofdstuk 6 over validiteit zal worden nagegaan of de gemeten concepten inderdaad overeenkomen met het begrip zoals bedoeld. Een geslaagde kalibratie op een unidimensionaal construct beschouwen we als een noodzakelijke voorwaarde voor begripsvaliditeit.

4.3.3 Normering en adviescategorieën

Op basis van de vastgestelde cumulatieve verdeling worden ruwe scores ('aantal goed') omgezet in percentielscores voor Taal, Rekenen-Wiskunde, Studievaardigheden, Wereldoriëntatie (indien van toepassing) en Totaal. Zowel ruwe scores als percentielscores worden vermeld op het zogenoemde leerlingrapport (zie figuur 4.3). Daarnaast wordt de totaalscore (zonder de scores te normaliseren, zoals dat in veel andere instrumenten wél gebeurt) omgezet in een standardscore met een bereik van 501 tot en met 550, een gemiddelde van (omstreeks) 535 en een standaarddeviatie van om en nabij de 9 (voor de werkelijke waarden, die van jaar tot jaar licht verschillen, zie tabel 6.2). De resulterende standardscores zijn op zichzelf tamelijk betekenisloos (ook al heeft het jarenlang omgaan met deze Eindtoetsscores in het onderwijs voor een duidelijk omschreven betekenis toekenning gezorgd). Hun werkelijke betekenis ontleent de getallen aan de adviescategorieën die eraan verbonden worden met betrekking tot de school- of brugklastypen die voor de betreffende leerling als het meest geschikt worden beschouwd.

Figuur 4.3 Voorbeeld van een leerlingrapport



NB. Dit betreft een voorbeeld; in werkelijkheid kunnen de gegevens die behoren bij een standardscore van '536' afwijken van de hier vermelde resultaten.

In de zogenoemde ‘poppetjesgrafiek’ wordt per brugklatype immers aangegeven hoeveel procent van de leerlingen met eenzelfde score in het betreffende brugklatype is geplaatst. Verder worden aan de leerkracht op grafische wijze ook aanvullende gegevens verstrekt over het percentage leerlingen dat met deze score in het betreffende brugklatype niet werd bevorderd naar het tweede leerjaar of werd overgeplaatst naar een lager brugklatype. Deze gegevens die ouders, leerkrachten en leerlingen steun moeten bieden bij het maken van een keuze, zijn ontleend aan de zogenoemde Toelatings- en doorstroomonderzoeken (T&DO) die Cito met regelmaat in samenwerking met het CBS organiseert. Omdat het T&DO in feite een follow- up-studie vormt, is het niet mogelijk om resultaten te presenteren voor de toetsversie van 2010. Vanwege de grote equivalentie van de verschillende toetsversies is dat ook niet nodig: de in deze verantwoording gerapporteerde resultaten kunnen van toepassing worden geacht op de toets van 2010. Het is om die reden ook overbodig om elk jaar zo’n kostbaar vervolgonderzoek te doen. Dat er om de twee à drie jaar toch zo’n onderzoek wordt gedaan, is met name om de maatschappelijke ontwikkelingen te kunnen volgen. Daarbij gaat het om mogelijke lichte verschuivingen in de doorstroom van leerlingen naar verschillende vormen van vervolgonderwijs.

Het meest recente T&DO betreft de leerlingen die in 2005 aan de Eindtoets Basisonderwijs hebben deelgenomen. De resultaten van 135 442 van de 153 728 deelnemende leerlingen (88,1%) zijn gekoppeld aan bij het CBS beschikbare individuele gegevens over het voortgezet onderwijs. Cito verkreeg op deze manier in het schooljaar 2005-2006 gegevens over de instroom van deze leerlingen in het eerste leerjaar van het voortgezet onderwijs. Daarvan werden 17 109 leerlingen (17,1%) niet meegenomen in verdere analyses omdat zij in het eerste leerjaar volgens het CBS in de brede brugklatypes ‘overig vmbo’ en ‘vmbo/avo’ of in het praktijkonderwijs werden geplaatst. In 2006-2007 en 2007-2008 werden daar respectievelijk de gegevens over de doorstroom naar het tweede en derde leerjaar aan toegevoegd. Ook konden eventuele doubleurs worden geïdentificeerd. Daarbij was het informatieverlies zeer gering: van slechts .6, respectievelijk .8% van de leerlingen in het bestand waren geen gegevens over het tweede, respectievelijk derde leerjaar bekend.

Bij de interpretatie van de score op de Eindtoets kan onderscheid gemaakt worden naar schooltype en naar brugklatype. Op basis van de doorstroomgegevens worden de intervallen vastgesteld die vermeld staan in tabel 4.5 en tabel 4.6. Wat betreft de brugklatypes in tabel 4.6 moet daaraan toegevoegd worden dat de grenzen worden bepaald door de standaardscores die noch op onderschatting, noch op overschatting van de mogelijkheden van de leerling lijken te wijzen (scores tussen het twintigste en het tachtigste percentiel per brugklatype).

Tabel 4.5 Interval dat het uitgangspunt vormt voor de interpretatie van de standaardscore per schooltype

Schooltype	Interval
basisberoepsgerichte leerweg (bb)	501-523
kaderberoepsgerichte leerweg (kb)	524-528
gemengde / theoretische leerweg (gt)	529-536
havo	537-544
vwo	545-550

Tabel 4.6 Interval dat het uitgangspunt vormt voor de interpretatie van de standaardscore per brugklatype

Brugklatype	Interval
basisberoepsgerichte leerweg	501-520
basis- en kaderberoepsgerichte leerweg	519-525
kaderberoepsgerichte leerweg	524-528
gemengde / theoretische leerweg	529-533
gemengde / theoretische leerweg en havo	532-536
gemengde / theoretische leerweg en havo / vwo	535-541
havo	537-540
havo / vwo	540-545
vwo	545-550

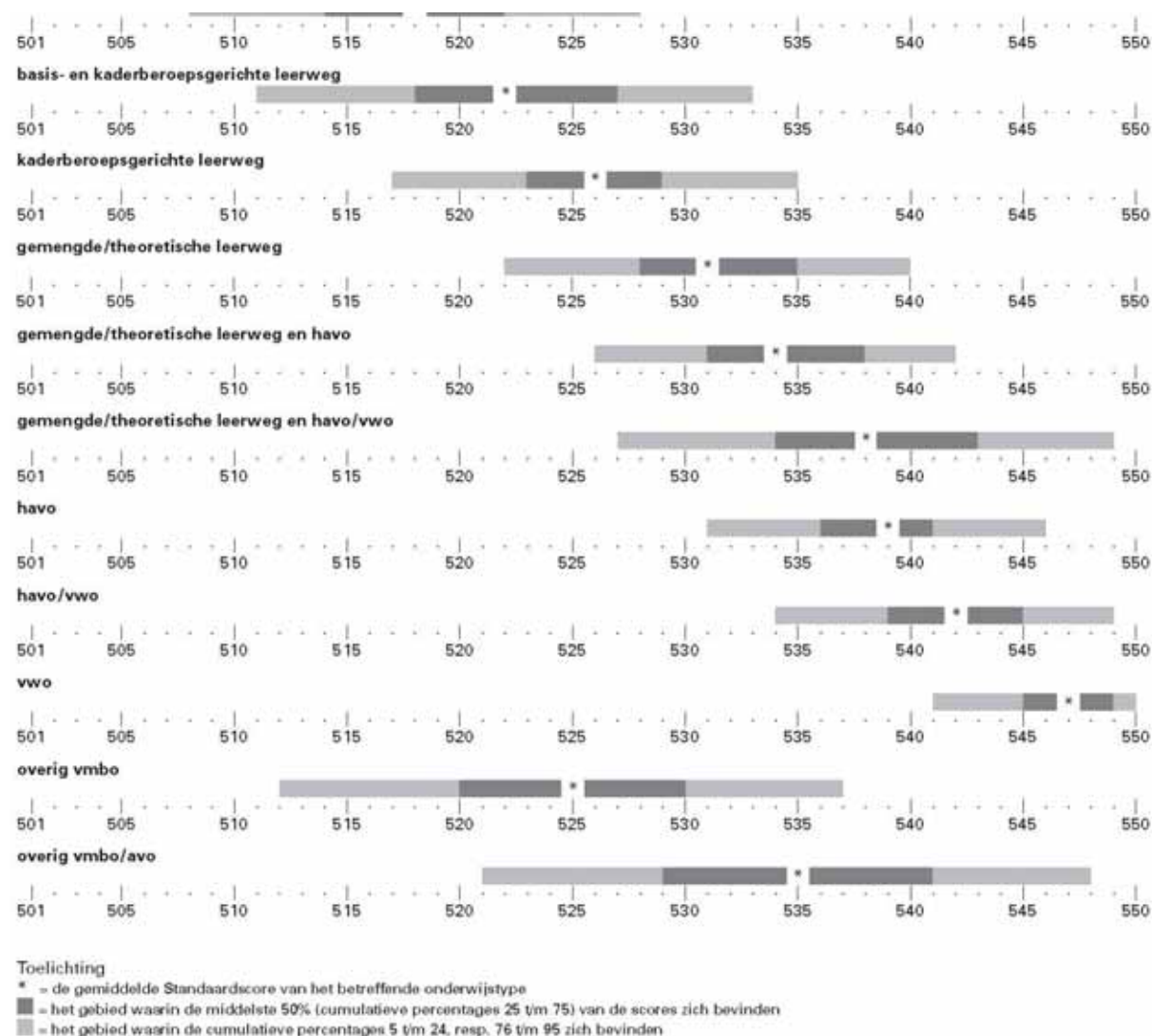
In figuur 4.4 is grafisch weergegeven hoe de prestatieverdeling voor de verschillende onderscheiden brugklastype er bij de instroom in leerjaar 1 in het schooljaar 2005-2006 uitzag. Per brugklastype is de gemiddelde score aangegeven en wordt door middel van balkjes zichtbaar gemaakt waar de grenzen liggen van de volgende drie standaardscoregebieden (in termen van percentages):

- 5% tot en met 24%
- 25% tot en met 75%
- 76% tot en met 95%

Het is eenvoudig vast te stellen dat de gemiddelde standaardscores systematisch oplopen naarmate het brugklastype een hoger niveau kent. Hetzelfde geldt voor de balkjes die de genoemde standaardscoregebieden representeren: deze schuiven systematisch verder naar rechts op. Dit geldt uiteraard niet voor de brede adviescategorieën 'overig vmbo' en 'vmbo / avo' die onderaan de figuur zijn toegevoegd (maar die voor de advisering niet gebruikt (kunnen) worden).

Meer resultaten van het Toelatings- en doorstroomonderzoek komen in hoofdstuk 6 bij de bespreking van de criteriumvaliditeit aan de orde.

Figuur 4.4 Gemiddelde en spreiding van standaardscores per brugklastype



5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Betrouwbaarheid

Voor het schatten van de betrouwbaarheid van de Eindtoets en de onderdelen daarvan is gebruik gemaakt van methoden uit zowel de klassieke testtheorie als de item respons theorie. We rapporteren betrouwbaarheden met betrekking tot de toetsonderdelen die ook daadwerkelijk worden gerapporteerd op het leerlingrapport (zie figuur 4.2). Dat zijn – naast de totaalscore – de scores op de onderdelen Taal, Rekenen-Wiskunde en Studievaardigheden. Ook voor het facultatieve onderdeel Wereldoriëntatie worden gegevens verstrekt.

Tabel 5.1 geeft de betrouwbaarheden in termen van alfa en de daarop gebaseerde 95%-betrouwbaarheidsintervallen.

Tabel 5.1 *Betrouwbaarheden en betrouwbaarheidsintervallen voor de EB 2010*

Onderdeel	Aantal opgaven	Alfa	MAcc	Test-hertest (simulatie)	95%-betrouwbaarheidsinterval (op basis van alfa)	
					Ruwe score	Standaardscore
Totaal	200	.95	.96	.96	±11.42	±4.10
Taal	100	.91	.91	.91	±7.45	
Rekenen-Wiskunde	60	.92	.92	.92	±6.03	
Studievaardigheden	40	.81	.82	.82	±4.54	
Wereldoriëntatie	90	.90	.91	.91	±3.62	

De alfa-coëfficiënten zijn zonder uitzondering hoog. Dat wil zeggen, voor de totaalscore geldt dat dit de score is op grond waarvan het advies ten aanzien van het best passende brugklatype wordt uitgebracht. Omdat aan dit advies zware consequenties voor leerlingen verbonden (kunnen) zijn, dienen aan de betreffende score de hoogste eisen te worden gesteld. Dat impliceert dat volgens de COTAN de betrouwbaarheidscoëfficiënt tenminste .90 moet bedragen om als goed gekwalificeerd te kunnen worden (Evers et al., 2010, p. 33). Aan de toetsonderdelen Taal, Rekenen-Wiskunde en Studievaardigheden hoeven deze zware eisen niet te worden gesteld omdat ze slechts bedoeld zijn om de sterke en zwakke punten van een leerling in kaart te brengen. Hetzelfde geldt voor het facultatieve en aanvullende onderdeel Wereldoriëntatie. Voor al deze onderdelen geldt dat een betrouwbaarheidscoëfficiënt ten minste .80 moet bedragen om als goed te kunnen worden aangeduid.

Geconcludeerd kan worden dat alle betrouwbaarheidscoëfficiënten als 'goed' kunnen worden benoemd.

Het is mogelijk om de betrouwbaarheid van de Eindtoets en de onderdelen daarvan te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de toetsen OPLM-geschaald zijn. De betrouwbaarheidsschattingen zijn berekend met behulp van het programma OPTAL (Verstralen, 1997).

In OPTAL wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met de betrouwbaarheidscoëfficiënt uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een toets, maar dan gezien als functie van de latente variabele θ . Deze verwachte waarde wordt aangeduid met $\tau(\theta)$. Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie wordt aangegeven met het symbool $\text{Var}(\tau)$. Tussen θ en $\tau(\theta)$ bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de

toetsscore $\tau(\theta)$ moet behalen (dat is alleen zo als de toets oneindig lang wordt). De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor met een eenmalige toetsafname niet meer zonder fout de waarde van θ bepaald kan worden. De variantie van de geobserveerde toetsscore wordt aangegeven met $\text{Var}(t|\tau(\theta))$, en door weer gebruik te maken van de distributie van θ in de populatie kan ook de gemiddelde variantie van de geobserveerde toetsscores berekend worden.

$$\text{Var}(t) = E[\text{Var}(t | \tau(\theta))] \quad (5.1)$$

Deze variantie kan opgevat worden als de (gemiddelde) meetfoutvariantie in de metriek van de geobserveerde scores t . In analogie met de theorie over de betrouwbaarheid volgt dan

$$\text{MAcc} = \frac{\text{Var}(\tau)}{\text{Var}(\tau) + \text{Var}(t)} \quad (5.2)$$

waarin MAcc staat voor 'Accuracy of Measurement'.

De hierboven beschreven werkwijze is alleen te volgen als een toets of toetsonderdeel in zijn geheel op een en dezelfde vaardigheidsdimensie is gekalibreerd. In de Eindtoets is dat niet overal het geval. Voor het onderdeel Taal bijvoorbeeld zijn de kalibraties in feite uitgevoerd op de onderdelen Begrijpend lezen, Schrijven, Spelling en Woordenschat als onderscheiden vaardigheidsdimensies. Om toch MAcc als schatting van de betrouwbaarheid te kunnen berekenen op het niveau van de toetsonderdelen die gerapporteerd worden in het leerlingrapport, is de kalibratie herhaald alsof de items steeds één onderliggende vaardigheidsdimensie voor de gehele toets, respectievelijk een toetsonderdeel representeren. Op deze manier is ook na te gaan hoe het gesteld is met de lokale betrouwbaarheid van de toets en de toetsonderdelen. In paragraaf 5.2 gaan we nader in op deze lokale betrouwbaarheid in termen van meetnauwkeurigheid.

De in tabel 5.1 weergegeven MAcc-betrouwbaarheidsschattingen vormen een vrijwel exacte afspiegeling van de overeenkomstige coëfficiënt alfa. Een enkele keer is MAcc een honderdste hoger, hetgeen overeenkomt met het feit dat alfa een lichte onderschatting vormt van de betrouwbaarheid (GLB's - Greatest Lower Bound – zijn in de regel bij deze hoogte van de coëfficiënt .01 of .02 hoger). Interpretatie van de waarden sluit dan ook volledig aan bij de conclusies die we hierboven hebben getrokken op basis van alfa.

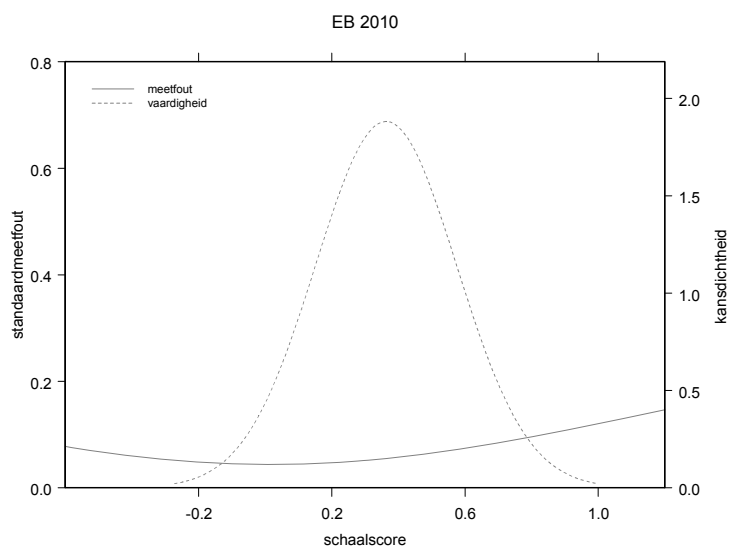
Er heeft geen test-hertestonderzoek plaatsgevonden. De afnamecontext van de Eindtoets leent zich hier niet goed voor. De toets ligt immers bij wijze van spreken op straat. De inhoud wordt uitgebreid besproken door de leerlingen onderling, met hun ouders en sommige opgaven worden zelfs besproken in het Jeugdjournaal. Het feit dat alle items echter OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd voor een groep van 150 000 leerlingen. Daarbij hebben we enerzijds de vaardigheidsverdeling van alle leerlingen in de Eindtoets 2010, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheidsscore aselekt uit de verdeling genomen en zijn twee bij deze vaardigheidsscore horende toetsafnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 150 000 dubbele (virtuele) afnames berekend. Men kan deze simulatie beschouwen als een test-hertestonderzoek onder ideale condities. De tweede toetsafname is immers volledig onafhankelijk van de eerste en wordt niet beïnvloed door de kennis die de leerling mogelijk verworven heeft via de eerste toetsafname. Daarnaast is er geen sprake van invloed van een test-hertestinterval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden. De resultaten zijn weergegeven in tabel 5.1. De uitkomsten komen wederom vrijwel exact overeen met eerder berekende coëfficiënten en leiden dan ook tot dezelfde conclusies met betrekking tot de betrouwbaarheid van de Eindtoets.

Ten slotte is in tabel 5.1 informatie opgenomen over de 95%-betrouwbaarheidsintervallen die horen bij de onderscheiden toetsscores. Ze zijn berekend op basis van alfa, maar gegeven de grote gelijkenis tussen de verschillende betrouwbaarheidsschattingen hadden we daar even goed de andere coëfficiënten voor kunnen gebruiken.

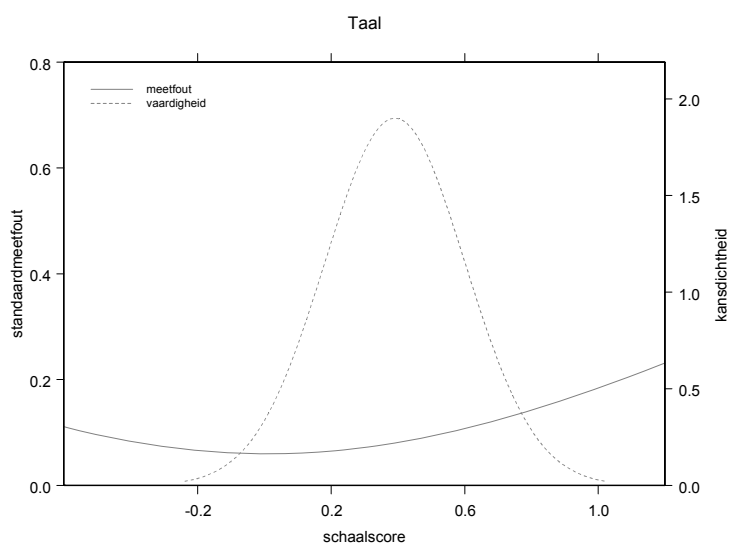
5.2 Nauwkeurigheid

De in tabel 5.1 vermelde betrouwbaarheidscoëfficiënten en betrouwbaarheidsintervallen hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen. De figuren 5.1 tot en met 5.5 geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de verschillende onderdelen en de totaalscore. In deze figuren waarin ook kansdichtheidfuncties zijn opgenomen, staat voor de score op de toets, c.q. ieder afzonderlijk toetsonderdeel, de grootte van de meetfout afgebeeld. Hierbij past de kanttekening dat de standaardmeetfout onder meer afhankelijk is van het aantal items in een onderdeel. Omdat de aantallen items per onderdeel verschillen, zijn de standaardmeetfouten in de figuren dus onderling niet direct vergelijkbaar. Daarnaast is de schaal waarop de kansdichtheid is aangegeven bij elke figuur steeds aangepast aan de gevonden waarden. Ook dit zorgt ervoor dat de figuren onderling niet overal rechtstreeks vergelijkbaar zijn.

Figuur 5.1 Grootte van de meetfout voor de totaalscore op de Eindtoets Basisonderwijs 2010 (inclusief de kansdichtheidfunctie voor deze score)

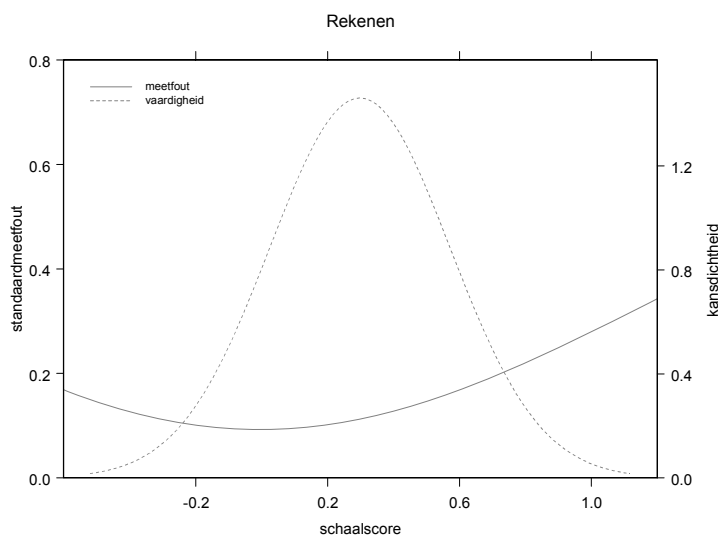


Figuur 5.2 Grootte van de meetfout voor de score op het onderdeel Taal van de Eindtoets Basisonderwijs 2010 (inclusief de kansdichtheidsfunctie voor deze score)

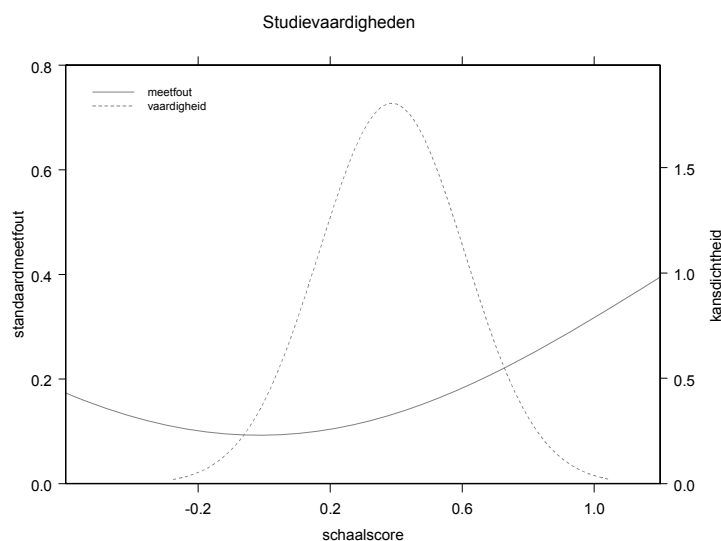


De figuren maken duidelijk dat de meetfout kleiner is in de lagere en gemiddelde vaardigheidsregionen dan in de hogere vaardigheidsregionen. Dit komt overeen met de bedoelingen van de toetsconstructeurs. Omdat in de lagere en gemiddelde vaardigheidsregionen tussen meer adviescategorieën moet worden gedifferentieerd dan in de hogere vaardigheidsregionen, is gekozen voor relatief gemakkelijke toetsen met een linksscheve verdeling. Het ligt voor de hand dat om die reden het differentiërend vermogen en de nauwkeurigheid in de lagere en gemiddelde regionen optimaal is.

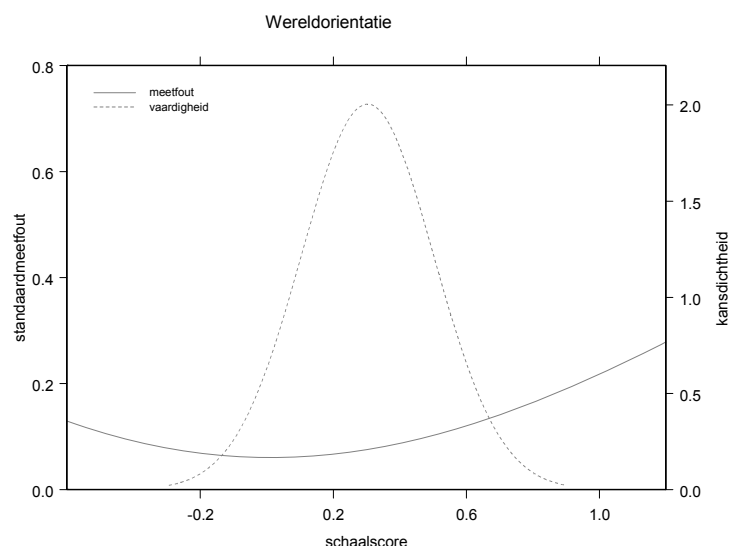
Figuur 5.3 Grootte van de meetfout voor de score op het onderdeel Rekenen van de Eindtoets Basisonderwijs 2010 (inclusief de kansdichtheidsfunctie voor deze score)



Figuur 5.4 *Grootte van de meetfout voor de score op het onderdeel Studievaardigheden van de Eindtoets Basisonderwijs 2010 (inclusief de kansdichtheidfunctie voor deze score)*



Figuur 5.5 *Grootte van de meetfout voor de score op het onderdeel Wereldoriëntatie van de Eindtoets Basisonderwijs 2010 (inclusief de kansdichtheidfunctie voor deze score)*



Het advies met betrekking tot het schooltype dat voor een leerling als het meest geschikt wordt aangemerkt wordt gebaseerd op de totaalscore (standaardscore). Daarom is het voor deze score interessant om na te gaan wat het effect van de hierboven beschreven betrouwbaarheid en nauwkeurigheid is op de verstrekte adviezen.

In tabel 5.2 is weergegeven welke percentages leerlingen op basis van hun werkelijke vaardigheidsscore een advies hebben gekregen dat overeenkomt met het advies op basis van de geschatte vaardigheidsscore. Voor de afkortingen met betrekking tot de schooltypen en de score-intervallen die horen bij iedere adviescategorie wordt verwezen naar tabel 4.5. In tabel 5.2 is bijvoorbeeld af te lezen dat 84.4% van de leerlingen met een havo-advies op basis van de geschatte vaardigheidsscore ook daadwerkelijk zo'n advies heeft gekregen op grond van zijn werkelijke vaardigheidsscore.

Tabel 5.2 Advies met betrekking tot best passend schooltype op basis van de werkelijke vaardigheidsscore als percentage van het advies op basis van de geschatte vaardigheidsscore

Adviescategorie (schooltype) waarin ware score valt	Adviescategorie met betrekking tot best passend schooltype				
	bb	kb	gt	havo	vwo
bb	85.7	11.1	.1	----	----
kb	14.0	62.2	6.6	----	----
gt	.3	26.7	77.7	7.3	----
havo	----	----	15.6	84.4	15.4
vwo	----	----	----	8.2	84.6
totaal	100	100	100	100	100

6 Validiteit

De primaire functie van de Eindtoets Basisonderwijs is het geven van onafhankelijke informatie voor de keuze van een passend brugklatype. De toets is geconstrueerd op basis van de verwachting dat hij iets kan zeggen over het onderwijs dat de leerling in de toekomst aankan. De Eindtoets is een school- of leervorderingentoets: de toets meet wat een leerling in vergelijking met andere leerlingen in acht jaar basisonderwijs geleerd heeft. En leervorderingen zeggen iets over de kansen op succes in de verschillende typen van het voortgezet onderwijs. Met andere woorden, leerprestaties in het verleden worden geacht iets te kunnen zeggen over leerprestaties in de toekomst. Deze claim zal empirisch worden onderbouwd in paragraaf 6.3 over criteriumvaliditeit. Daar zullen we met name ingaan op een specifieke vorm van criteriumvaliditeit, namelijk predictieve validiteit 'over tijd'. In deze paragraaf staat de rapportage van het zogenoemde Toelatings- en doorstroomonderzoek centraal.

Daaraan vooraf gaat paragraaf 6.1 over inhoudsvaliditeit. We kunnen ons daar beperken tot het formuleren van onze conclusies. De inhoud van de Eindtoets is immers in hoofdstuk 3 al uitgebreid aan de orde geweest.

In paragraaf 6.2 komt de vraag aan de orde of de Eindtoets meet wat de toets beoogt te meten.

De begripsvaliditeit van de Eindtoets wordt hier via verschillende invalshoeken benaderd en empirisch onderbouwd. Daarbij wordt tevens aandacht besteed aan de equivalentie van de Eindtoets 2010 ten opzichte van eerdere versies van de Eindtoets.

6.1 Inhoudsvaliditeit

De inhoudsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de opgaven in een toets een welomschreven en afgebakend universum representeren van mogelijk in de toets op te nemen opgaven. De opgaven in de Eindtoets sluiten nauw aan bij het doel en de inhoud van het onderwijs op de basisschool. Dat moet ook, want zij dienen een goede afspiegeling te vormen van wat leerlingen in het basisonderwijs in acht jaar tijd geleerd hebben. Daarbij moeten het opgaven zijn waarin leerstof behandeld wordt die elke basisschoolleerling zou kunnen beheersen, uitgaande van de communale onderwijsdoelen en -inhouden. Voor een optimaal voorspellend vermogen is dit noodzakelijk.

Een uitgebreide inhoudelijke analyse van de inhoud van de Eindtoets staat in hoofdstuk 3 van deze verantwoording. Hieronder vatten we nog eens samen hoe Cito de toetsinhoud representatief maakt voor de communale onderwijsdoelen en -inhouden en hoe structuur en inhoud van de toets van jaar tot jaar constant worden gehouden.

- De Eindtoets Basisonderwijs heeft een vaste structuur met drie subtoetsen die betrekking hebben op Taal, Rekenen-Wiskunde en Studievaardigheden. Daarnaast is er het (facultatieve) onderdeel Wereldoriëntatie. Binnen deze hoofdgebieden zijn weer rubrieken onderscheidbaar. Binnen het onderdeel Taal wordt onderscheid gemaakt tussen Begrijpend lezen, Schrijven, Spelling en Woordenschat. De drie onderdelen van Rekenen-Wiskunde (Getallen en bewerkingen; Verhoudingen, breuken en procenten; Meten, meetkunde, tijd en geld) correleren onderling hoger dan .97 en worden als één homogene schaal opgevat. Het onderdeel Studievaardigheden valt uiteen in Studieteksten, Informatiebronnen, Schema's, tabellen en grafieken en Kaartlezen. Bij Wereldoriëntatie worden Aardrijkskunde, Geschiedenis en Natuuronderwijs onderscheiden.
- De toetsonderdelen kennen vaste aantallen opgaven. De verdeling van de aantallen opgaven over de subtoetsen en binnen Wereldoriëntatie vormt een afspiegeling van de onderwijsbesteding in het reguliere basisonderwijs met betrekking tot de onderscheiden vakgebieden. Daarnaast laten de intercorrelaties tussen de toetsonderdelen een patroon zien dat van jaar tot jaar praktisch constant is en wordt gehouden. Adequate equivaleringsprocedures vormen de basis van deze constante toetsopbouw.
- Cito analyseert voortdurend doelstellingen en methoden en gaat aan de hand van de resultaten na of de Eindtoets nog voldoende aansluit bij de inhoud van het onderwijs.

- Cito publiceert een inhoudsverantwoording, het Doelenboek (Staphorsius, 2003, 2009) en stelt deze zo nodig bij. In het Doelenboek staan doelstellingen en voorbeeldopgaven. Alle deelnemende scholen ontvangen dit Doelenboek, waardoor de toetsinhoud transparant wordt gemaakt voor ieder die daarin is geïnteresseerd. Een samenvatting van het Doelenboek zoals dat maatgevend is geweest voor de inhoud van de Eindtoets 2010 is besproken in hoofdstuk 3. Uit het Doelenboek en de samenvatting valt af te leiden dat de toetsinhoud minutieus wordt voorgeschreven aan de toetsconstructeurs.
- Een onderdeel van de jaarlijkse proeftoets is het verzoek om commentaar van leerkrachten op gehanteerde onderwijsdoelstellingen en opgaven. Ook anderszins is de Eindtoets een 'glazen huis': iedereen kan (na afname) kennisnemen van de toetsinhoud en deze desgewenst becommentariëren. De afgelopen jaren heeft de toetsinhoud (als representatie van communale onderwijsdoelen) niet ter discussie gestaan.
- De constructie van opgaven vindt plaats in commissies waarvan de leden ervaring hebben in het basisonderwijs.

Op basis van de beschreven procedures en bevindingen kan geconcludeerd worden dat – uitgaande van communale onderwijsdoelen en -inhouden – de inhoudsvaliditeit van de Eindtoets uitstekend en vrijwel onomstreden is.

6.2 Begripsvaliditeit

De begripsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de toetsscores toe te schrijven zijn aan de concepten en constructen die deel uitmaken van het theoretische kader dat aan de ontwikkeling van de toets ten grondslag ligt. Hoewel het gebruik van de termen concepten en constructen in het kader van leervorderingentoetsen misschien wat lastig is, zijn er ook bij instrumenten als de Eindtoets Basisonderwijs verschillende manieren om empirische evidentie te verzamelen die kan worden opgevat als onderbouwend voor de begripsvaliditeit ervan. We zullen deze hieronder achtereenvolgens bespreken.

6.2.1 Passing van het meetmodel; psychometrische kwaliteit van de opgaven

De opgaven die geconstrueerd worden met als doel te worden opgenomen in de Eindtoets, vormen na proeftoetsing en kalibratie een gekalibreerde opgavenbank. Bij de analyse van de antwoorden van de leerlingen op de opgaven wordt nagegaan of de verschillende opgaven en opgaventypen een beroep doen op hetzelfde complex aan vaardigheden (zie paragraaf 4.3.1). Items die niet voldoen aan de beschreven passingscriteria worden uit de verzameling verwijderd.

In paragraaf 4.3.2 is beschreven hoe het gehanteerde meetmodel is getoetst en zijn gegevens gepresenteerd waaruit af te leiden valt of en in welke mate de kalibratie geslaagd genoemd kan worden. Wij vatten hier de conclusies samen. Uit de S-toetsen en grafische weergaven daarvan bij een toetsonderdeel blijkt dat de afzonderlijke opgaven steeds één onderliggende vaardigheidsdimensie representeren. Dit wordt bevestigd door de uitstekende R^2 -waarden en de nauwkeurigheid van de itemparameterschattingen. Voor elk van de toetsonderdelen waarvoor de assumptie van unidimensionaliteit geldt, kan dus worden geconcludeerd dat er sprake is van een unidimensionaal concept (respectievelijk latente trek of vaardigheidsscore) onder de opgaven in het desbetreffende toetsonderdeel. In termen van klassieke testtheorie sluiten de uitkomsten van de itemanalyses (de R^2 -waarden) en hoge interne consistentie (Cronbach's alfa) hierbij aan.

Deze uitkomsten zijn op zichzelf nog geen bewijs voor begripsvaliditeit (meet de toets wat hij geacht wordt te meten?), maar vormen daarvoor wel een noodzakelijke voorwaarde als eerste stap in de bewijsvoering.

6.2.2 Structuur

Eerder is aangegeven dat de Eindtoets een vaste opbouw en structuur kent die tot uitdrukking komt in een relatief constante intercorrelatiematrix voor de toetsonderdelen (zie hiervoor paragraaf 2.4.2.1 over equivaleringsprocedures en paragraaf 6.1 over inhoudsvaliditeit). In tabel 2.1 is deze matrix bij wijze van illustratie opgenomen voor de Eindtoetsversies van 2004 en 2005. Om de continuïteit van de toets ten

opzichte van eerdere jaren te controleren, is in tabel 6.1 een soortgelijke matrix opgenomen voor de meest recente versies van 2009 (boven de diagonaal) en 2010 (onder de diagonaal). Zowel de structuurkenmerken voor de versie van 2010 (dat is nu eenmaal de toetsversie die hier – in engere zin – verantwoord wordt), als de continuïteit in deze structuur zijn vanuit het oogpunt van begripsvaliditeit belangrijk.

Het patroon in de correlaties voor de jaren 2009 en 2010 lijkt nog steeds heel erg op elkaar, zoals dat ook al kon worden vastgesteld voor de jaren 2004 en 2005. De onderlinge verschillen tussen de correlaties voor het ene en voor het andere jaar zijn soms nihil (voor Studievaardigheden en Begrijpend lezen) of zeer klein en ten hoogste .053 (voor Woordenschat en Studievaardigheden; vetgedrukt in de tabel). Het gemiddelde verschil tussen de correlaties is slechts .025. Vergelijken we het patroon voor de twee jaren die het verst uit elkaar liggen (2004 en 2010), dan mag de conclusie luiden dat er sprake is van een stabiele situatie, ook al is het gemiddelde verschil tussen de correlaties iets hoger, namelijk .039. Het hoogste verschil in correlatie voor beide jaren is .077.

Tabel 6.1 Latente correlaties tussen onderdelen van de EB (2009 boven de diagonaal, ter vergelijking 2010 onder de diagonaal; vergelijk ook tabel 2.1)

	Rekenen-Wiskunde	Begrijpend lezen	Schrijven	Spelling	Woordenschat	Studievaardigheden
Rekenen-Wiskunde	1	.666	.672	.555	.592	.837
Begrijpend lezen	.607	1	.913	.607	.862	.896
Schrijven	.677	.944	1	.690	.764	.900
Spelling	.522	.587	.652	1	.539	.611
Woordenschat	.539	.818	.753	.492	1	.783
Studievaardigheden	.834	.896	.903	.563	.752	1

Wat valt er over dit patroon inhoudelijk op te merken? Op de eerste plaats is uit de intercorrelaties (uitgaande van de toetsversie van 2010) af te lezen, dat Spelling een vrij unieke plaats inneemt. De vier laagste intercorrelaties zijn immers correlaties met Spelling en alle vijf correlaties met Spelling behoren tot de laagste zes intercorrelaties. Verder valt op dat de hoogste correlatie betrekking heeft op Schrijven en Begrijpend lezen (.944); in alle besproken toetsversies was dit het geval met correlaties van steeds boven de .90. Ook Woordenschat laat met beide onderdelen een (binnen de taalopgaven) relatief sterke samenhang zien, die terug te voeren is op het semantische karakter van de opgaven (dit in tegenstelling tot het niet-semantische karakter van Spelling). De correlaties met Studievaardigheden ten slotte laten een terugkerend patroon zien: hoge tot zeer hoge correlaties met Begrijpend lezen en Schrijven (om en nabij de .90), gevolgd door nog altijd hoge correlaties met achtereenvolgens Rekenen-Wiskunde en Woordenschat, met Spelling als hekkensluiter.

Zoals al eerder aangegeven correleren de drie onderdelen van Rekenen-Wiskunde (Getallen en bewerkingen; Verhoudingen, breuken en procenten; Meten, meetkunde, tijd en geld) onderling hoger dan .97. Ook binnen Studievaardigheden zijn de correlaties tussen de onderdelen hoog (variërend tussen .79 en .89 in 2010), zij het iets minder hoog dan bij Rekenen-Wiskunde. De onderdelen van de facultatieve toets Wereldoriëntatie laten nog iets minder samenhang zien: Natuuronderwijs en Aardrijkskunde correleren in 2010 .74, Natuuronderwijs en Geschiedenis .78 en Aardrijkskunde en Geschiedenis .84. Dat komt overeen met het feit dat deze onderdelen aspecten van wereldoriëntatie kunnen worden genoemd, maar in de praktijk toch dikwijls afzonderlijke vakgebieden vormen.

6.2.3 Equivalentie met eerdere toetsversies

De hierboven besproken intercorrelatiematrices illustreren de (constante) structuur van de onderdelen in de Eindtoets, maar vormen tegelijkertijd ook bewijsmateriaal voor de equivalentie van de 2009- en 2010-versie van de Eindtoets ten opzichte van eerdere versies. Deze informatie kan worden aangevuld met gegevens over de omvang van de statistische fout die bij het equivaleren onvermijdelijk wordt gemaakt, de equivaleringsfout. Het betreft soortgelijke gegevens als de gegevens die eerder bij wijze van illustratie in tabel 2.2 werden gepresenteerd voor de versies van 2004 tot 2006 (zie paragraaf 2.4.2.1). Wat betreft de toegepaste procedures kan worden opgemerkt dat deze ten opzichte van de aangegeven jaargangen nog verder zijn vervolmaakt. In 2009 en 2010 heeft de equivalering namelijk niet meer plaatsgevonden op basis van de proeftoetsing, maar op basis van de procedure van de stabiele scholen. Dat wil zeggen op basis van 'sample 0', een subgroep van scholen die zowel qua deelname aan de Eindtoets, als qua aantal deelnemende leerlingen en gemiddelde toetsprestaties een grote continuïteit laten zien (voor een nadere beschrijving van de procedure zie paragraaf 2.4.2.1). Vergelijkende analyses hebben laten zien dat de equivalering op basis van stabiele scholen nog net iets betere resultaten laat zien. De equivalering op basis van proeftoetsing wordt nog steeds uitgevoerd, maar nu bij wijze van controle. Daarbij is het aantal observaties in de proeftoetsing verdubbeld; elke opgave wordt minimaal twee keer in een proeftoets opgenomen, waardoor de N is verdubbeld. Ook de resultaten op basis van sample 1, sample 2, sample 3 en de samengevoegde samples 1+2+3 worden beschouwd als controle op de uitkomsten. Dan is er nog de equivalering op basis van ankeropgaven. Deze procedure is ten opzichte van 2006 aanzienlijk verbeterd. Het aantal ankeropgaven bedraagt inmiddels 30 à 35. Deze worden opgenomen in de Eindtoets van een aselecte groep leerlingen (N=1200 in 2008; N=3200 in 2009 en N=2400 in 2010) naast de 165 à 170 opgaven die ook bij alle overige leerlingen worden afgenomen. Door het opnemen van ankeropgaven in de Eindtoets is een alternatieve equivalering mogelijk; ook de resultaten hiervan gelden als controle. Kortom, de equivalering vindt plaats op grond van sample 0 uit de 'stabiele scholen equivalering', terwijl controle op de uitkomsten plaatsvindt op grond van de andere samples enerzijds en IRT analyses op de proeftoetsing en via de ankeropgaven anderzijds.

*Tabel 6.2 Vergelijking van uitkomsten van de equivaleringsprocedures voor de Eindtoets 2010
(uitkomsten voor 2008 en 2009 ter vergelijking)*

	parameter	2008	2009	2010
sample 0	gemiddelde	535.35	535.46	535.85
	SD	9.64	9.46	9.35
sample 1	gemiddelde	535.30	535.27	535.61
	SD	9.57	9.62	9.45
sample 2	gemiddelde	535.46	535.28	535.49
	SD	9.63	9.58	9.43
sample 3	gemiddelde	535.22	535.38	535.43
	SD	9.72	9.50	9.42
sample 1+2+3	gemiddelde	535.33	535.31	535.52
	SD	9.64	9.57	9.39
op basis van proeftoets	gemiddelde	536.16	536.15	535.87
	SD	9.60	9.67	8.92
op basis van ankeropgaven	gemiddelde	536.07	535.38	535.98
	SD	9.59	9.97	8.90

Alle resultaten ten aanzien van gemiddelde en standaardafwijking zijn vermeld in tabel 6.2, waarin de 'officiële' equivalering vetgedrukt is.

De overeenkomsten zoals deze tot uitdrukking komen in de tabel zijn net als die voor de jaren 2004, 2005 en 2006 (zie tabel 2.3) zeer goed te noemen. Dat geldt voor de onderlinge vergelijking van de 2009- en de 2010-versie, maar ook voor de vergelijking met de eerder in tabel 2.3 gepresenteerde cijfers. De kwaliteit van de toegepaste procedures is dus voor 2009 en 2010 onveranderd hoog gebleven, waarbij nog kwaliteitswinst is geboekt door aanpassing en uitbreiding van de (controle)procedures.

De uitkomsten van de analyses die in deze en de vorige paragraaf (én eerder in paragraaf 2.4.2.1) zijn gerapporteerd, maken duidelijk dat de verschillende versies van de Eindtoets van jaar tot jaar hun structuur behouden in termen van intercorrelaties tussen de onderdelen. Daarmee wordt het construct dat in de Eindtoets wordt geoperationaliseerd constant gehouden. De equivaleringsprocedures zorgen er daarnaast voor dat de gemiddelde score en variantie daarin van jaar tot jaar vrijwel exact vergelijkbaar blijven. Dit geldt voor de Eindtoets 2010 net zo duidelijk als voor eerdere versies van de toets. Belangrijk is dat deze resultaten zijn op te vatten als onderbouwing van de begripsvaliditeit van de toets. Equivalentie is op deze wijze immers te beschouwen als een bepaalde vorm van soortgenootvaliditeit (in termen van gelijkenis tussen qua concrete inhoud volledig verschillende toetsen die operationalisaties vormen van hetzelfde construct). Het kan constructie-technisch gesproken opgevat worden als een prestatie van formaat dat een dergelijke equivalentie tot stand wordt gebracht. In het toetsconstructieproces kan in principe immers van alles mis gaan wat afbreuk doet aan die equivalentie. Als er bijvoorbeeld van jaar tot jaar steeds aanmerkelijke verschillen zouden worden vastgesteld in de intercorrelatiematrix of de gemiddelde moeilijkheidsgraad, zou dit in hoge mate afbreuk hebben gedaan aan de validiteit van de Eindtoets. Het feit dat dit niet gebeurt wijst erop dat de toegepaste constructieprocedures in orde zijn en dat er geen sprake is van toevallige of systematische fouten in die procedures die de validiteit zouden kunnen bedreigen.

Andere gegevens met betrekking tot soortgenootvaliditeit komen aan de orde in paragraaf 6.2.4. We wijzen er nogmaals op dat door de vérgaande equivalentie van de toetsversies de resultaten van eerder verricht valideringsonderzoek ook van toepassing zijn op de Eindtoets 2010.

6.2.4 Soortgenootvaliditeit; convergente en discriminante validiteit

Met betrekking tot de convergente en discriminante validiteit verkeert Cito in de gelukkige omstandigheden dat veel leerlingen die de Eindtoets maken een jaar eerder in groep 7 ook hebben deelgenomen aan de Entreetoets. De Entreetoets kent in grote lijnen dezelfde opbouw in onderdelen als de Eindtoets (net als bij de Eindtoets Basisonderwijs worden Rekenen, Taal en Studievaardigheden gemeten) en is dus, hoewel de toets een jaar eerder wordt afgenomen, een echte soortgenoot op het gebied van leervorderingen. Daarbij willen we benadrukken dat de Eindtoets en de Entreetoets verre van identiek zijn. Er zijn aanmerkelijke verschillen tussen beide toetsen: geen enkele opgave is gelijk, de wijze van afname en het afnametijdstip verschillen en de scoreverdelingen wijken af (waardoor het ook geen parallelle toets is).

Vanwege de gelijkenis in opbouw is het mogelijk een analyse te maken van de convergente en discriminante validiteit van de toetsonderdelen. Daarbij moet echter wel worden aangetekend dat de samenstelling van het onderdeel Studievaardigheden in de Eindtoets aanzienlijk afwijkt van de opgaven in de Entreetoets. Daarom gaat de gelijkenis tussen beide toetsen voor dit onderdeel aanmerkelijk minder goed op dan voor de onderdelen Taal en Rekenen-Wiskunde.

In tabel 6.3 zijn de correlaties gegeven voor de totaalscores op de beide toetsen en voor de toetsonderdelen. De analyses hebben betrekking op een dataset van 107 001 leerlingen; vanwege ontbrekende gegevens is de minimale N iets lager, namelijk 106 580.

Tabel 6.3 Correlaties tussen Eindtoets 2010 en Entreetoets 2009

		Entreetoets 2009		
	Taal	Rekenen- Wiskunde	Studie- vaardigheden	Totaal
Eindtoets 2010				
Taal	.831	.602	.711	.801
Rekenen-Wiskunde	.605	.839	.678	.760
Studievaardigheden	.581	.482	.556	.595
Totaal	.808	.786	.785	.871

De correlatie tussen beide totaalscores is zeer hoog, namelijk .871. Dat is een waarde die in een test-hertestonderzoek naar de betrouwbaarheid niet zou misstaan. Daarbij moet men zich realiseren dat de Eindtoets is afgenomen in februari 2010 en de Entreetoets in de regel tussen april en juli 2009 (op sommige scholen in de eerste weken van het schooljaar 2009-2010 in groep 8). De tussenliggende periode omvat dus circa vijf à tien maanden. En verder betreft het natuurlijk ook niet hetzelfde instrument; van de opgaven voor beide toetsen is er geen een hetzelfde. Kortom, de soortgenootvaliditeit van de Eindtoets als geheel is zeer groot. Bovendien wordt hiermee de kwaliteit van de constructie-, kalibratie- en equivaleringsprocedures die we eerder in deze verantwoording al uitgebreid hebben beschreven en geëvalueerd nog eens extra onderstreept.

Aan de hoogte van de correlaties is te zien dat Studievaardigheden een vreemde eend in de bijt is. De oorzaak daarvan is al genoemd: de overeenkomst tussen de opgaven is niet erg groot. Voor Taal en Rekenen-Wiskunde is die overeenkomst er wel. De betreffende correlaties laten een goed herkenbaar patroon zien, met steeds de hoogste correlatie op de (grijs gearceerde) diagonaal voor de overeenkomstige toetsonderdelen. Er is dus sprake van convergente en discriminante validiteit. Daarbij is het duidelijk dat ook de correlaties met inhoudelijk verschillende onderdelen hoog zijn: in leervorderingen is er sprake van een grote gemeenschappelijke variantie als het gaat om de onderscheiden leerinhouden en leerstofgebieden.

Een soortgelijke correlatieanalyse is ook uit te voeren op de verschillende onderscheiden aspecten van respectievelijk Taal en Rekenen-Wiskunde. Studievaardigheden moeten we hierbij buiten beschouwing laten (zie boven). De resultaten staan in tabel 6.4.

Voor de onderdelen Taal geldt dat de hoogste correlaties steeds op de diagonaal staan. Daarbij hebben we de beide onderdelen Spelling (niet-werkwoorden versus werkwoorden) opgevat als behorende tot hetzelfde leerstofdomein. Er is goed te onderkennen dat het enerzijds gaat om onderdelen waarbij de betekenis van taal (semantiek) een belangrijke rol spelen (Schrijven, Begrijpend lezen en Woordenschat) en anderzijds om onderdelen waarbij dat niet het geval is (Spelling). De correlaties tussen de 'semantische' onderdelen zijn hoog (tussen .53 en .72), die tussen 'semantische' en 'niet-semantische' onderdelen aanzienlijk lager (tussen .28 en .49).

Tabel 6.4 Correlaties tussen Eindtoets 2010 en Entreetoets 2009 voor de onderdelen van respectievelijk Taal en Rekenen-Wiskunde

Eindtoets 2010	Entreetoets 2009				
Onderdelen Taal					
	Schrijven	Spelling niet-werkwoorden	Spelling werkwoorden	Begrijpend lezen	Woordenschat
Schrijven	.713	.492	.431	.694	.606
Spelling niet-werkwoorden	.370	.609	.487	.346	.280
Spelling werkwoorden	.413	.524	.471	.338	.332
Begrijpend lezen	.698	.452	.391	.716	.655
Woordenschat	.529	.342	.275	.569	.697
Onderdelen Rekenen-Wiskunde					
	Getallen en bewerkingen	Verhoudingen, breuken en procenten	Meten, meetkunde, tijd en geld		
Getallen en bewerkingen	.742	.713	.740		
Verhoudingen, breuken en procenten	.728	.744	.753		
Meten, meetkunde, tijd en geld	.676	.689	.719		

De onderdelen van Rekenen-Wiskunde laten geen discriminante validiteit zien. Dat is ook niet te verwachten aangezien de intercorrelaties tussen deze onderdelen binnen de Eindtoets (en ook voor de Entreetoets) zeer hoog zijn. Eerder hebben we al aangegeven dat deze intercorrelaties voor de Eindtoets hoger zijn dan .97. Alle onderdelen en opgaven van Rekenen-Wiskunde hebben dus betrekking op dezelfde vaardigheidsdimensie. Correlaties tussen deze onderdelen van de Eindtoets enerzijds en de Entreetoets anderzijds liggen tussen .68 en .75 en verschillen dus erg weinig van elkaar.

6.2.5 De Eindtoets over de jaren heen en in relatie tot diverse subgroepen

In deze paragraaf wordt een aantal uitkomsten van analyses besproken die eerder zijn gepubliceerd in de *Terugblik en resultaten 2010* (Cito, 2010). Dit is een publicatie die jaarlijks wordt uitgebracht. De uitkomsten die interessant zijn voor de begripsvaliditeit van de Eindtoets vatten we hieronder samen. Zij vormen – in termen van de onderbouwing van de externe structuur van de toets – een aanvulling op het onderzoek naar de convergente en discriminante validiteit dat we in de vorige paragraaf hebben gerapporteerd. Het betreft onderzoek naar verschillen tussen relevante groepen en samenhangen met relevante factoren. Daarbij beperken we ons tot variabelen die deel uitmaken van het nomologische netwerk rondom het construct leervorderingen zoals dat door de Eindtoets wordt gemeten, variabelen waarvoor het mogelijk was theoretisch gefundeerde verwachtingen te formuleren. Als voorbeelden noemen we het leerlinggewicht, speciale deelnemersgroepen, mate van verstedelijking (grote steden versus overige gemeenten) en leeftijd (vertraagd versus niet-vertraagd). Empirische bevestiging van dergelijke verwachtingen vormt, zoals bekend, een ondersteuning van de begripsvaliditeit. Hetzelfde geldt voor de samenhangen met intelligentie die in paragraaf 6.3 aan de orde komen.

Trend 2005-2010 landelijk en voor de vier grote steden (G4)

Deze sectie laat zien hoe de gemiddelde standardscore op de Cito Eindtoets zich 'gedraagt' over meerdere jaren bezien. Daarbij gaat het enerzijds om het 'landelijke' beeld, dat wil zeggen alle deelnemende leerlingen in alle deelnemende scholen. Anderzijds gaat het om het totaal van de G4, dat is de groep van vier grote steden (Amsterdam, Utrecht, Rotterdam, Den Haag). Over de G4 wordt gerapporteerd omdat de samenstelling van de leerlinggroepen (bijvoorbeeld naar leerlinggewicht) in deze steden veelal afwijkend is van het landelijke beeld. Voornamelijk vanwege dit leerlinggewicht is te verwachten dat de gemiddelden voor de G4 lager zullen uitvallen dan de landelijke gemiddelden (de afgelopen jaren is dit ook steeds het geval geweest).

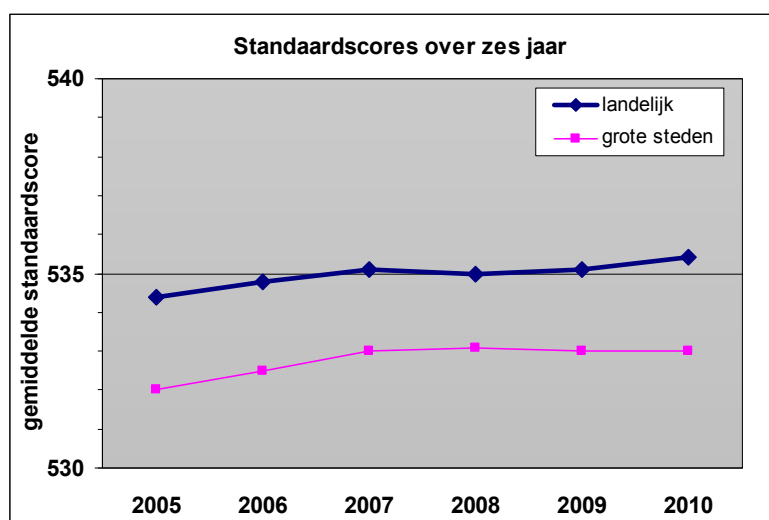
Als we de resultaten van de vier grote steden vergelijken met de landelijke resultaten moeten we ons realiseren dat het hierbij niet gaat om een vergelijking van twee verschillende groepen, maar om de vergelijking tussen een kleine groep en een grote groep waarbij de kleine groep deel uitmaakt van de grote groep. De leerlingen van de vier grote steden maken immers deel uit van de landelijke groep leerlingen. Dit kan tot gevolg hebben dat sommige verschillen enigszins gedempt worden.

In tabel 6.5 en figuur 6.1 rapporteren we de landelijke gemiddelde standardscores¹⁰ en het gemiddelde van de vier grote steden in de periode 2005 t/m 2010. Het landelijk gemiddelde is de afgelopen zes jaar vrij stabiel te noemen. De trend die landelijk zichtbaar is, zien we ook terug bij de G4. Ook hier is de gemiddelde standardscore over de jaren heen stabiel.

Tabel 6.5 Verdeling gemiddelde standardscores (landelijk en voor de G4) van 2005-2010

	2005	2006	2007	2008	2009	2010
Landelijk	534.4	534.8	535.1	535.0	535.1	535.4
G4-totaal	532.0	532.5	533.0	533.1	533.0	533.0

Figuur 6.1 Standardscores over zes jaar landelijk en voor de G4



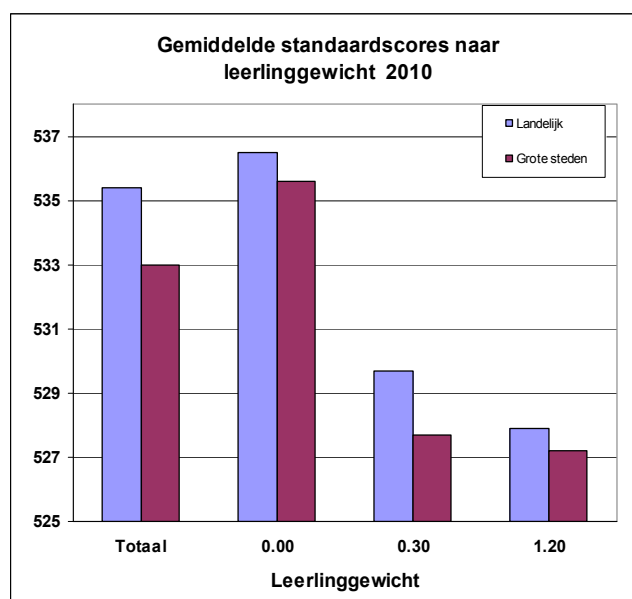
¹⁰ De hier gerapporteerde gemiddelden zijn gemiddelde leerlinggemiddelden: dit zijn gemiddelden gebaseerd op alle deelnemende leerlingen als eenheid van analyse. In de schoolrapportage bij de Eindtoets worden de scores van deelnemende scholen vergeleken met gemiddelde schoolgemiddelden: dit zijn gemiddelden gebaseerd op alle deelnemende scholen als eenheid van analyse. Hierdoor kunnen er kleine verschillen optreden tussen de op verschillende plekken gerapporteerde gemiddelden.

Leerlinggewicht

Op het antwoordblad van de Eindtoets kan het leerlinggewicht worden aangestreept. Leerlinggewichten worden toegekend in het kader van het Formatiebesluit WPO (Wet Primair Onderwijs). Dat besluit regelt de formatie voor bestrijding van onderwijsachterstanden. Voor een toelichting op het begrip leerlinggewicht verwijzen we naar hoofdstuk 4.

In figuur 6.2 zijn de gemiddelde standaardscores van de naar leerlinggewicht onderscheiden groepen afgebeeld voor het jaar 2010. In tabel 6.6 geven we de grootte van de onderscheiden groepen weer in aantallen en percentages samen met de gemiddelde standaardscore van de onderscheiden groepen in 2010. Leerlingen bij wie geen leerlinggewicht is aangestreept of bij wie meer dan één leerlinggewicht is aangestreept, zijn in de tabel opgenomen in de kolom met het vraagteken. Deze leerlingen zijn niet opgenomen in figuur 6.2.

Figuur 6.2 Gemiddelde standaardscores 2010 naar leerlinggewicht landelijk en voor de G4



In tabel 6.6 zien we dat de vier grote steden in vergelijking met het landelijk beeld relatief meer leerlingen met een hoger leerlinggewicht herbergen. Voor de vier grote steden geldt dat in 2010 22.0% van de leerlingen een 1.20-leerling is, terwijl dat in het hele land 6.0% is. De totale groep 1.20-leerlingen bestaat voor 43.7% uit leerlingen in de vier grote steden. Verder zien we dat leerlingen met een hoger leerlinggewicht zoals verwacht gemiddeld een lagere standaardscore behalen. Wel moet opgemerkt worden dat er voor een groot aantal leerlingen (landelijk gezien 15.5%) geen (eenduidig) leerlinggewicht op het antwoordblad is ingevuld waardoor de resultaten met voorzichtigheid geïnterpreteerd dienen te worden. Eerder zagen we dat leerlingen in de vier grote steden gemiddeld duidelijk lager presteren dan het landelijk gemiddelde. Oorzaak daarvan is dat de samenstelling van de leerlinggroepen naar leerlinggewicht in de vier grote steden verschilt van het landelijke beeld. In de grote steden behoort een groot aantal leerlingen (ongeveer een derde) tot de 0.30- en 1.20-leerlingen. De gemiddelden in deze groepen zijn duidelijk lager dan het gemiddelde van de groep 0.00-leerlingen.

Tabel 6.6 Aantal, percentage leerlingen en gemiddelde standaardscore 2010 landelijk en voor de G4 uitgesplitst naar leerlinggewicht

2010 – Landelijk	Totaal	0.00	0.30	1.20	?
Aantallen	147100	104136	11420	8758	22786
Percentages	100	70.8	7.8	6.0	15.5
Standaardscore	535.4	536.5	529.7	527.9	536.6

2010 - Grote steden	Totaal	0.00	0.30	1.20	?
Aantallen	17365	10070	1761	3825	1709
Percentages	100	58.0	10.1	22.0	9.8
Standaardscore	533.0	535.6	527.7	527.2	535.9

Speciale deelnemersgroepen

Op het antwoordblad van de Eindtoets kan een speciale code worden aangestreept (de 'leerlingcategorie') met behulp waarvan leerlingen kunnen worden ingedeeld in speciale deelnemersgroepen¹¹. We herhalen hier de omschrijving van deze codes zoals deze ook op bladzijde 18 en 19 van de Handleiding 2010 staat.

- **Code I**
(Allochtone) leerlingen die aan het begin van groep 8 vier jaar of korter in Nederland zijn en die het Nederlands onvoldoende beheersen om de opgaven in de Eindtoets Basisonderwijs 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).

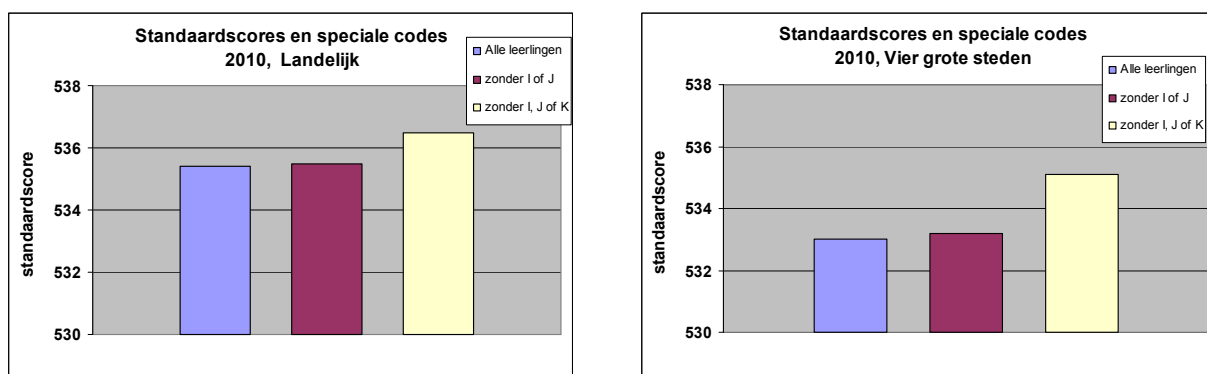
Het is ook mogelijk dat er twee codes bij eenzelfde leerling worden aangestreept (I en J of I en K). Omdat er zeer weinig leerlingen met zo'n dubbele code zijn, zijn deze leerlingen in de totalen opgenomen als leerlingen met code I.

Met betrekking tot de deelname van deze speciale groepen aan de Eindtoets geldt dat leerlingen met een code I of J buiten de doelgroep vallen waarvoor de Eindtoets is bedoeld. Dit wordt in de handleiding ook expliciet zo gesteld (zie bladzijde 7 van de handleiding). Voor leerlingen met een code K wordt aangegeven dat de toets in principe geschikt is, maar dat deze voor een deel van de leerlingen aan de moeilijke kant zal zijn. Als alternatief voor deze leerlingen wordt de Niveautoets (zie bladzijde 2) geopperd omdat deze qua moeilijkheid beter afgestemd is op het niveau van leerlingen met een leerachterstand. Hoe scholen of schoolbesturen in de praktijk met deze doelgroepdefinitie omgaan, met andere woorden in welke mate zij leerlingen van de speciale deelnemersgroepen al dan niet aan de Eindtoets laten deelnemen, is informatie die bij Cito niet bekend is. Wat uiteraard bij Cito wel bekend is, is het aantal leerlingen met een code-aanduiding dat wel aan de Eindtoets deelneemt. In 2010 had 0.40% van de leerlingen een I-code, 0.53% een J-code en 5.71% een K-code; 93.4% had geen code. De meeste leerlingen met een speciale code behoren tot de 1.20 leerlingen.

¹¹ Niet voor alle leerlingen is er op het antwoordblad een code aangestreept. Ook de juistheid van de aangestreepte codes is door Cito niet of slechts beperkt controleerbaar. Cito beschikt niet over andere middelen om deze achtergrondkenmerken in kaart te brengen. Vandaar dat wij ons in deze wetenschappelijke verantwoording en in alle andere wetenschappelijke verslagen van Cito op de op het antwoordblad ingevulde gegevens baseren.

Tabel 6.5 gaf een overzicht van de gemiddelde standaardscores. Deze gemiddelden zijn gebaseerd op alle leerlingen die aan de Eindtoets Basisonderwijs hebben deelgenomen. In figuur 6.3 zijn deze gemiddelde standaardscores opnieuw (grafisch) weergegeven, maar nu alleen voor 2010 en aangevuld met gegevens gebaseerd op alle leerlingen **exclusief de leerlingen met een I- of J-code en exclusief de leerlingen met een I-, J- of K-code** (dus voor alle leerlingen waarvoor geen code werd aangestreept). Dit zowel landelijk als voor de G4 afzonderlijk. Met name bij de gemiddelde standaardscores exclusief I-, J- en K-leerlingen valt op dat deze aanzienlijk hoger liggen dan de gemiddelde standaardscores exclusief I- en J-leerlingen en de gemiddelde standaardscores gebaseerd op alle deelnemende leerlingen. Dit is overeenkomstig de verwachting dat leerlingen met een speciale code, waarvan mag worden aangenomen dat zij problemen zullen ervaren in het volgen van het reguliere basisonderwijs, de gemiddelden in negatieve zin zullen beïnvloeden. Dit effect is voor de G4 (verschil met het overall gemiddelde is 2.1) groter dan landelijk (verschil is 1.1).

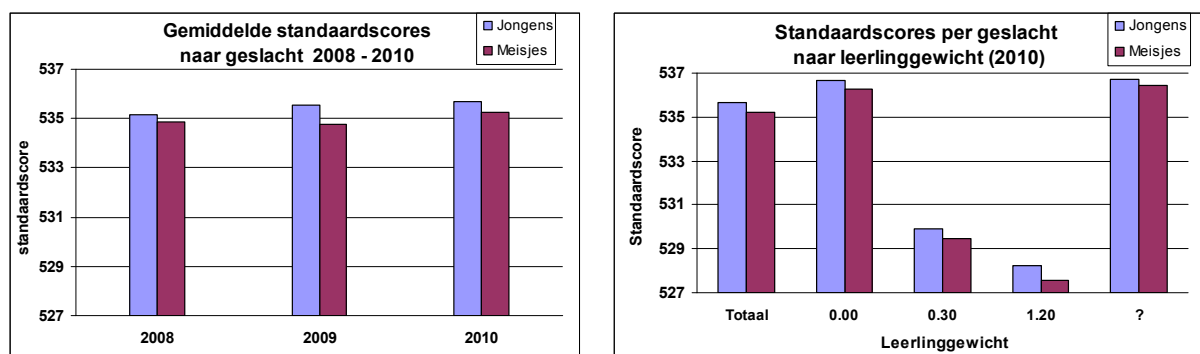
Figuur 6.3 Gemiddelde standaardscores gebaseerd op alle leerlingen, alle leerlingen exclusief I of J en alle leerlingen exclusief I, J of K



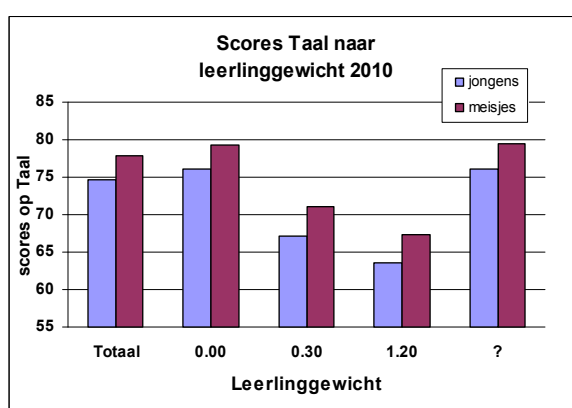
Sekse

In het linker paneel van figuur 6.4 worden voor 2008, 2009 en 2010 de gemiddelde standaardscores van de aan de Eindtoets deelnemende jongens en meisjes met elkaar vergeleken. In alle jaren behalen jongens gemiddeld iets hogere standaardscores dan meisjes. In 2008 scoorden de jongens gemiddeld 0.3 standaardscorepunt hoger dan de meisjes, in 2009 bedroeg dit verschil 0.8 standaardscorepunt en in 2010 is het verschil 0.4 standaardscorepunt. Het verschil tussen jongens en meisjes is in 2010 dus licht gedaald ten opzichte van 2009. Bij de bespreking van de afzonderlijke onderdelen wordt deze afname verder toegelicht. In het rechter paneel van figuur 6.4 worden deze gemiddelden uitgesplitst naar leerlinggewicht. De jongens behalen in alle gewichtscategorieën gemiddeld nét iets hogere standaardscores dan de meisjes. We kunnen in deze uitkomsten ook differentiëren naar toetsonderdeel (Taal, Rekenen-Wiskunde en Studievaardigheden). Figuur 6.5 laat zien dat het verschil tussen jongens en meisjes wat betreft de standaardscores niet verklaard kan worden uit de taalscores (aantal goed van 100 opgaven). De meisjes behalen gemiddeld hogere taalscores dan de jongens en dat verschil is er in alle gewichtscategorieën. Het gemiddelde verschil (voor alle leerlingen) is 3.3 scorepunten. Dit verschil bedroeg ook in 2008 3.3 scorepunten en in 2009 2.0 scorepunten.

Figuur 6.4 Gemiddelde standaardscores naar geslacht in 2008-2010 en voor 2010 uitgesplitst naar leerlinggewicht

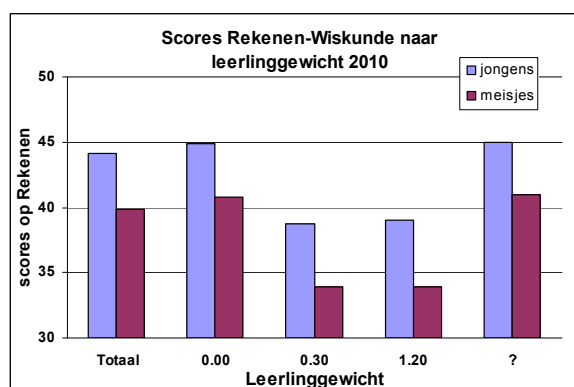


Figuur 6.5 Scores Taal naar geslacht en leerlinggewicht voor 2010



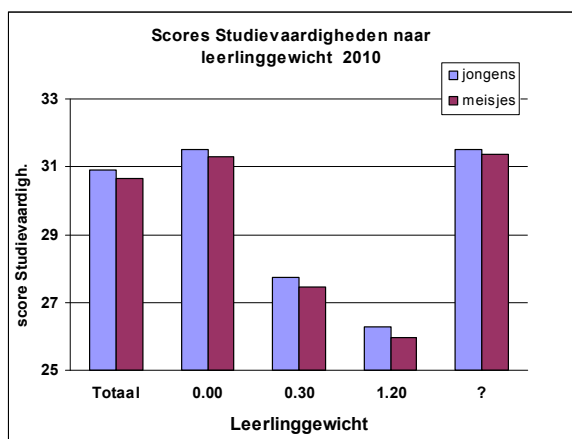
In de figuren 6.6 en 6.7 geven we soortgelijke afbeeldingen voor de onderdelen Rekenen-Wiskunde en Studievaardigheden. Bij Rekenen-Wiskunde behalen de jongens gemiddeld een hogere score dan de meisjes. Het verschil is echter groter dan bij Taal: in 2010 behalen de jongens gemiddeld 4.2 scorepunten meer voor Rekenen-Wiskunde dan de meisjes en ook hier geldt dat het verschil in alle gewichtscategorieën in het voordeel van de jongens is. In 2008 behaalden de jongens gemiddeld 3.9 scorepunten meer dan de meisjes en in 2009 was dat 4.2 scorepunten (telkens op een maximum van 60 punten).

Figuur 6.6 Scores Rekenen-Wiskunde naar geslacht en leerlinggewicht voor 2010



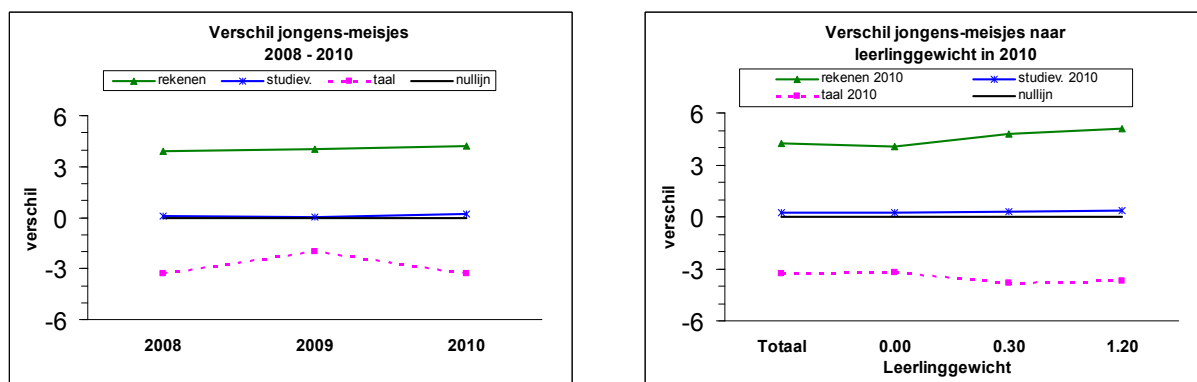
Bij het onderdeel Studievaardigheden doen de jongens het eveneens beter dan de meisjes en ook hier is dit verschil in alle gewichtscategorieën aanwezig. Het verschil is echter zeer gering en lijkt over de tijd heen vrij stabiel (tussen 0.05 en 0.2 punten op een totaal van 40 opgaven).

Figuur 6.7 Scores Studievaardigheden naar geslacht en leerlinggewicht voor 2010



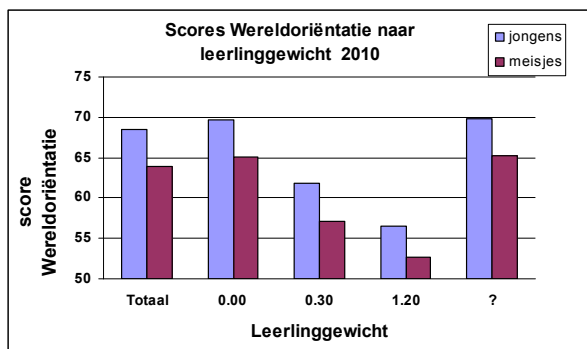
In het linker paneel van figuur 6.8 zijn voor de jaren 2008, 2009 en 2010 de verschillen tussen jongens en meisjes voor de drie onderdelen van de Eindtoets die meetellen voor de berekening van de standaardscore (Taal, Rekenen-Wiskunde en Studievaardigheden) nog eens samengevat. Verschillen boven de nullijn duiden op een verschil ten voordele van de jongens (Rekenen en Studievaardigheden); verschillen onder de nullijn (Taal) geven het verschil aan ten voordele van de meisjes. In het rechter paneel staan de verschillen tussen jongens en meisjes op de Eindtoets 2010 uitgesplitst naar leerlinggewicht. Samenvattend kunnen we zeggen dat jongens en meisjes op de verschillende onderdelen (gemiddeld genomen) verschillend presteren en dat dit geldt voor alle gewichtscategorieën.

Figuur 6.8 Verschillen tussen jongens en meisjes in 2008-2010 en voor 2010 uitgesplitst naar leerlinggewicht



In figuur 6.9 staat een overzicht waaruit het verschil tussen jongens en meisjes op het gebied van (het facultatieve onderdeel) Wereldoriëntatie kan worden afgelezen. De jongens halen op het onderdeel Wereldoriëntatie gemiddeld hogere scores dan de meisjes en dat geldt voor alle gewichtscategorieën. Het gemiddelde verschil tussen jongens en meisjes is 5.0 punten in 2008, 5.0 punten in 2009 en 4.6 punten in 2010.

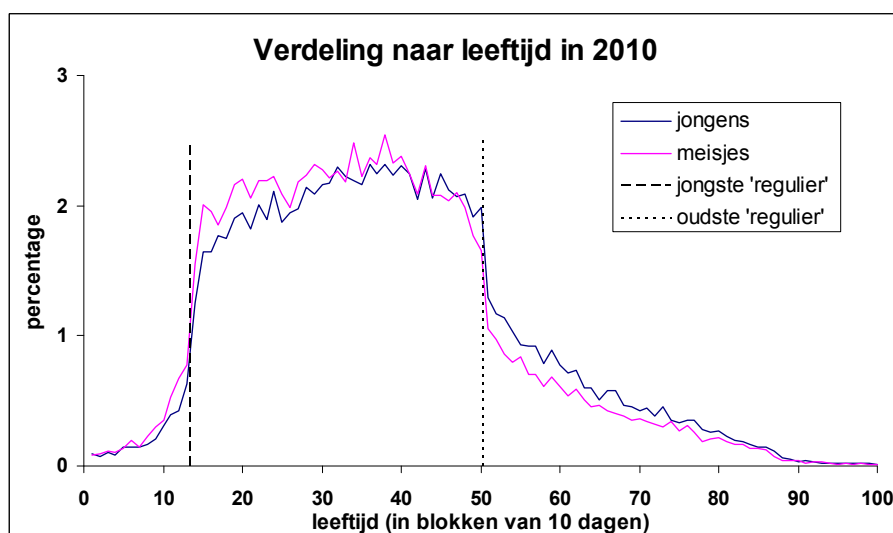
Figuur 6.9 Scores Wereldoriëntatie naar geslacht en leerlinggewicht voor 2010



Verschillen naar leeftijd

Niet alle leerlingen in groep 8 zijn even oud. Zijn er effecten van de verschillen in leeftijd op de hoogte van de scores op de Eindtoets Basisonderwijs? In figuur 6.10 staat de verdeling naar leeftijd van de leerlingen die in 2010 hebben deelgenomen aan de toets, voor jongens en meisjes afzonderlijk. De leeftijd op de horizontale as is uitgedrukt in eenheden van 10 dagen (een punt komt overeen met een periode van 10 dagen). Volgens de regelgeving die gold voor de invoering van het Basisonderwijs (begin jaren '80) werden leerlingen pas tot de eerste klas toegelaten als ze zes jaar werden vóór 1 oktober van het lopende jaar. Als die regel nu nog zou worden toegepast, zou dit betekenen dat leerlingen die zonder vertraging hun schoolloopbaan volbrengen en in februari 2010 in groep 8 zitten, geboren zijn tussen 1 oktober 1997 en 30 september 1998. De twee verticale stippellijnen in de figuur komen met deze data overeen en bakenen dus een periode van exact één jaar af. Uit de scherpe stijging en de scherpe daling van de twee curves rond die plaatsen kunnen we concluderen dat deze regel in de praktijk nog steeds wordt toegepast. Leerlingen met een code groter dan 50 (rechts van de rechtse stippellijn) zijn leerlingen die om een of andere reden vertraging hebben opgelopen, ze zijn bijvoorbeeld later begonnen in groep 3 of zijn gedoubleerd. Leerlingen met een leeftijdscode kleiner dan 14 (links van de linkse stippellijn) zijn 'voorlijke' leerlingen die waarschijnlijk eerder aan groep drie begonnen zijn dan volgens de oude regelgeving toegestaan was of die een klas hebben overgeslagen.

Figuur 6.10 Verdeling van aan de Eindtoets 2010 deelnemende leerlingen naar leeftijd

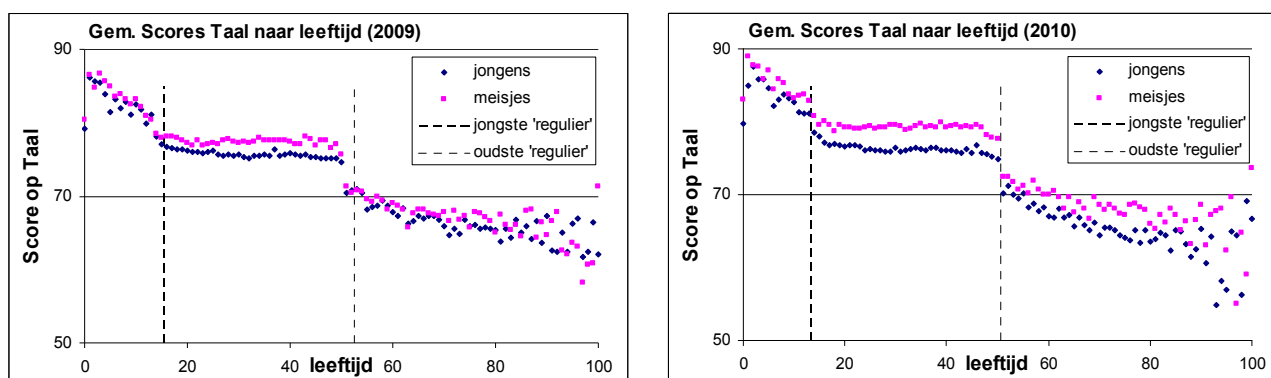


Er zijn opmerkelijke zaken op te merken aan de hand van figuur 6.10. Ten eerste: er zijn meer voorlijke meisjes dan jongens. Ten tweede zijn er bij de vertraagde leerlingen meer jongens dan meisjes. Tot slot zijn bij de reguliere leerlingen de jongens gemiddeld ouder dan de meisjes, want bij de jongsten onder de regulieren zijn de meisjes in de meerderheid.

In de figuren 6.11 tot en met 6.14 zijn achtereenvolgens per leeftijdsgroep de verschillen tussen jongens en meisjes op het gebied van de onderdelen Taal, Rekenen-Wiskunde, Studievaardigheden en Wereldoriëntatie weergegeven. Het 'uitwaai'-effect aan de uiteinden van de plotlijnen heeft te maken met het aantal observaties: dat is daar tamelijk gering. Het linker gedeelte van elke figuur heeft steeds betrekking op 2009, het rechter gedeelte op 2010. In alle figuren kunnen we zien dat de groep vertraagde leerlingen gemiddeld duidelijk lager scoort dan de leerlingen die normaal zijn doorgestroomd. En op alle onderdelen scoren de 'voorlijke' leerlingen hoger dan de overige leerlingen.

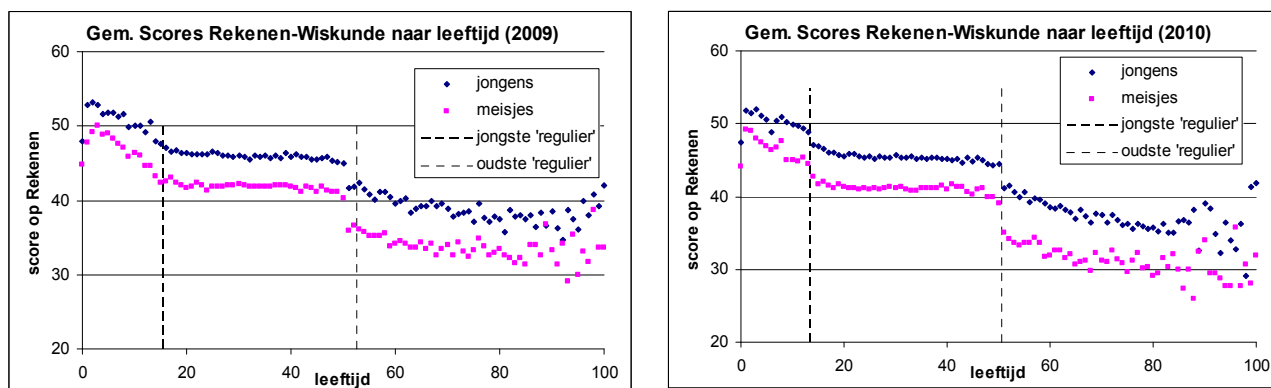
In figuur 6.11 zijn per leeftijdsgroep de verschillen tussen jongens en meisjes op het gebied van het onderdeel Taal weergegeven. De verschillen tussen jongens en meisjes zijn in 2010 in alle leeftijdsgroepen toegenomen ten opzichte van 2009. We wijzen nogmaals op een mogelijk foutieve interpretatie die zou kunnen volgen uit een vergelijking van de twee gedeeltes van de figuur. Bij de reguliere groep (tussen de twee stippellijnen) is duidelijk te zien dat de scores van de jongens in 2010 hoger zijn dan in 2009. Dit betekent echter niet automatisch dat de taalvaardigheid in 2010 is toegenomen; het verschil zou ook kunnen zijn veroorzaakt door een verschil in moeilijkheid van de opgaven of de onvergelijkbaarheid van de deelnemende groepen.

Figuur 6.11 Scores Taal naar leeftijd en geslacht voor 2009 en 2010



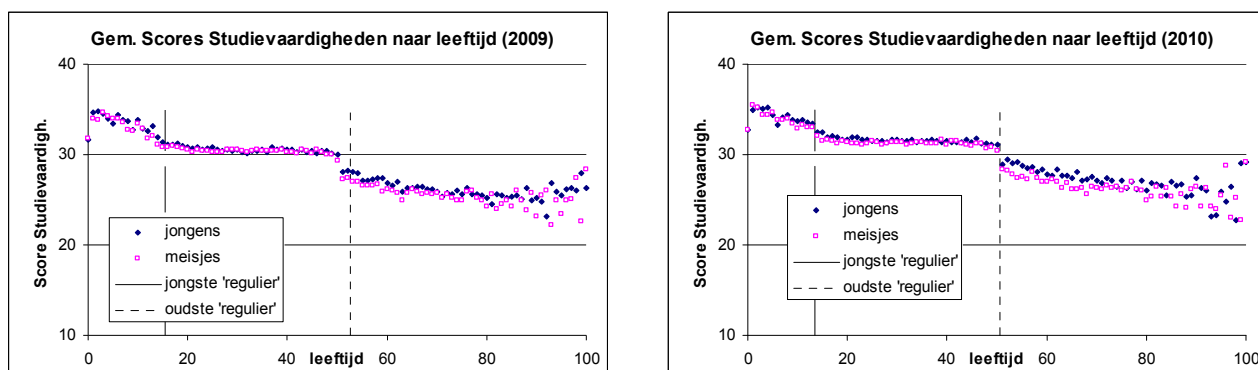
In figuur 6.12 staan per leeftijdsgroep de resultaten van jongens en meisjes op het onderdeel Rekenen-Wiskunde. De verschillen tussen jongens en meisjes zijn voor alle leeftijdsgroepen duidelijk en zijn in 2010 in alle leeftijdsgroepen ongeveer gelijk aan 2009.

Figuur 6.12 Scores Rekenen-Wiskunde naar leeftijd en geslacht voor 2009 en 2010



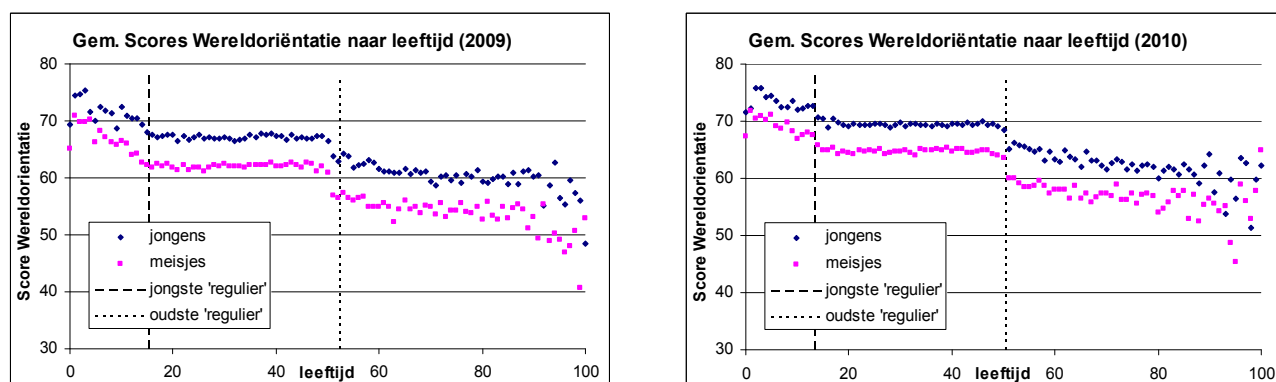
In figuur 6.13 zijn per leeftijdsgroep de verschillen tussen jongens en meisjes op het gebied van het onderdeel Studievaardigheden weergegeven. Er zijn in alle leeftijdsgroepen nauwelijks verschillen tussen jongens en meisjes zowel in 2009 als in 2010.

Figuur 6.13 Scores Studievaardigheden naar leeftijd en geslacht voor 2009 en 2010



In figuur 6.14 zijn per leeftijdsgroep de verschillen tussen jongens en meisjes op het gebied van het onderdeel Wereldoriëntatie weergegeven. De verschillen tussen jongens en meisjes zijn in 2010 in alle leeftijdsgroepen enigszins afgenomen ten opzichte van 2009.

Figuur 6.14 Scores Wereldoriëntatie naar leeftijd en geslacht voor 2009 en 2010



6.2.6 De Eindtoets en allochtone leerlingen; DIF-analyses

Eerder onderzoek naar DIF en itembias

Sinds de invoering van de Eindtoets is het aantal allochtone leerlingen in het basisonderwijs voortdurend toegenomen. Daarmee is de vraag hoe leerlingen met een verschillende herkomst en thuistaal het op de Eindtoets 'doen' niet alleen relevant, maar ook steeds urgenter geworden. Enerzijds is immers bekend dat allochtone leerlingen nog steeds een structurele achterstand laten zien ten opzichte van autochtone leerlingen (dit ondanks een geleidelijke verbetering van hun onderwijsresultaten; zie Van Schilt-Mol, 2007), anderzijds is het maatschappelijk belang dat aan de Eindtoets Basisonderwijs wordt gehecht groot. De bruikbaarheid van de toets voor verschillende doelgroepen heeft daarom grote implicaties voor de (onderwijs)toekomst van grote groepen leerlingen. Uiterwijk (1994) en Van Schilt-Mol (2007) hebben dit probleem geformuleerd in termen van 'Differential Item Functioning' (DIF). Daarmee doelen zij op het verschijnsel dat toetsen opgaven kunnen bevatten die verschillend 'functioneren' voor verschillende subgroepen leerlingen, zelfs wanneer deze leerlingen een vergelijkbaar prestatieniveau laten zien. In de IRT houdt dit concreet in dat leerlingen die behoren tot verschillende categorieën (bijvoorbeeld jongens versus meisjes, Nederlands sprekend versus anderstalig) dezelfde kans moeten hebben om op een item met een bepaalde respons te reageren. Er is sprake van DIF (ook wel aangeduid als 'onbedoelde moeilijkheden') als leerlingen uit verschillende subgroepen met een vergelijkbaar prestatieniveau een ongelijke kans hebben om een toetsopgave juist te beantwoorden. Wanneer de oorzaak daarvan niet behoort tot het construct dat de opgave beoogt te meten, is er sprake van itembias.

Uit eerder onderzoek (Uiterwijk, 1994; Uiterwijk & Vallen, 1997, 2005) is gebleken dat de Eindtoets (in jaren voorafgaand aan deze publicaties) opgaven bevatte waarbij sprake is van DIF en waarbij mogelijk sprake is

van itembias. Van Schilt-Mol (2007) heeft nauwkeurig onderzoek verricht naar DIF en itembias in de Eindtoets van 1997. In de Eindtoets Basisonderwijs van 1997 bleken 32 van de 240 opgaven DIF te vertonen, waarvan er 21 in het nadeel en 11 in het voordeel van de groep Turkse en Marokkaanse leerlingen bleken uit te pakken. Het merendeel hiervan bevond zich in de toetsonderdelen Taal en Wereldoriëntatie. Bij 9 van deze 32 opgaven was er sprake van bias in die zin, dat de betreffende moeilijkheid niet terug te voeren was op de vaardigheid die de opgave beoogde te meten. Voor nog 5 andere opgaven werd herziening wenselijk geacht. Van deze 14 opgaven droegen er 10 bij aan de score waarop het Citoadvies is gebaseerd (4 hadden betrekking op het facultatieve Wereldoriëntatie). Het totale effect van deze 10 opgaven met onbedoelde moeilijkheden bedroeg naar schatting minder dan 1 standaardscorepunt. Het was daarom gerechtvaardigd te concluderen dat de Cito Eindtoets Basisonderwijs leerlingen van allochtone afkomst niet benadeelt en ook niet bevoordeelt, althans voor zover het de onderzochte versie betreft.

Ondanks deze (voor de versie van 1997) gunstige uitkomsten heeft Cito inmiddels een aantal opgaventypen aangepast. Het onderzoek van Van Schilt-Mol bevatte daartoe een aantal bruikbare aanwijzingen en aanbevelingen. Ook werd het onderzoek waarop de selectie van opgaven is gebaseerd zodanig uitgebreid dat het nog beter mogelijk is de selectie van opgaven met onbedoelde moeilijkheden te voorkomen.

Differential Item Functioning in de Eindtoetsversie van 2010

Op de 200 opgaven van de Eindtoets (2010) waarop de standaardscore is gebaseerd zijn DIF-analyses uitgevoerd die vergelijkbaar zijn met de analyses van Van Schilt-Mol (2007) (zie hierboven). Probleem was wel dat door de afwijkende definitie van leerlinggewichten geen eenduidige informatie meer beschikbaar is over de etnische herkomst van leerlingen. Bij de afname van de Eindtoets wordt standaard wél informatie verzameld over de thuistaal van leerlingen, die overigens in lang niet alle gevallen wordt verstrekt.

Besloten is om DIF-analyses uit te voeren naar leerlinggewicht en thuistaal. Aanvullend werden ook DIF-onderzoek naar sekse gedaan. Voor de analyses naar leerlinggewicht werd een aselechte steekproef getrokken uit de data, voor elk van de drie categorieën 1200 leerlingen. Bij de variabele thuistaal bestond de steekproef uit 1000 leerlingen uit de grote groep met alleen Nederlands als thuistaal. De categorieën met een afwijkende thuistaal in de analyse werden gevormd door alle leerlingen die hadden aangegeven dat zij thuis respectievelijk alleen Turks (1100 leerlingen), Arabisch (1450), Surinaams (350), een andere West-Europese taal dan Nederlands (600) of nog een andere taal (1850) spraken. Omdat de software slechts vier groepen toelaat in de analyse, werden de onderscheiden categorieën op zinvolle wijze samengevoegd.

Voor elk van de vele honderden vergelijkingen werd de Mantel Haenszel DIF-statistiek berekend en op significantie getoetst door middel van de bijbehorende z-waarde. De resultaten voor de versie van 2010 zijn opgenomen in tabel 6.7.

Tabel 6.7 Resultaten van DIF-analyses voor thuistaal, leerlinggewicht en sekse voor 200 opgaven van de Eindtoets versie 2010 (exclusief Wereldoriëntatie)

Aantal items met significante DIF-resultaten				
	Aantal opgaven	Thuistaal	Leerlinggewicht	Sekse
Taal	100	7	7	7
Rekenen-Wiskunde	60	3	0	8
Studievaardigheden	40	2	0	2

Bezien we de resultaten voor thuistaal en leerlinggewicht, dan blijkt voor Rekenen-Wiskunde en Studievaardigheden het aantal significante toetsingsresultaten gelijk aan of lager te zijn dan wat men op basis van het gehanteerde kansniveau (.05) zou mogen verwachten. Ook voor Taal is het aantal significante toetsingsresultaten (met zeven per variabele) nauwelijks hoger dan verwacht. Daarbij moet bovendien in aanmerking worden genomen dat de steekproef een behoorlijke omvang kent, zodat de toetsingen relatief snel significant uitpakken. Nadere inspectie van de S-curven per leerlingcategorie en van

de opgaven zelf in het geval van een significant toetsingsresultaat leverde dan ook geen inhoudelijk interpreteerbare inzichten op, hetgeen onderstreept dat het hier zeer waarschijnlijk om 'toevalligheden' gaat. Ten slotte is via inspectie van gemiddelde moeilijkheidsgraden per categorie nagegaan hoe groot het effect is van de vastgestelde verschillen op de uiteindelijke toetsscore. Dit effect is in alle gevallen zeer minimaal. Kortom, het aantal significante resultaten is verwaarloosbaar te noemen, inhoudelijk niet interpreteerbaar en het effect op de score is vrijwel nihil. Verder kan men constateren dat het aantal opgaven met DIF in vergelijking met het onderzoek van Van Schilt-Mol (2007) is afgenomen van 32 op 240 naar maximaal 12 op 200.

Voor sekse leveren de DIF-analyses soortgelijke resultaten op. Wél is het aantal significante toetsingen (8) voor Rekenen-Wiskunde wat groter dan men mag verwachten op basis van kanskapitalisatie. Zes verschillen zijn in het voordeel van meisjes, twee in het voordeel van jongens. Ook hier leveren de aanvullende inspecties op S-curven en inhoud echter geen interpreteerbare resultaten op en zijn de effecten op de score minimaal.

De DIF-analyses (achteraf) gaven geen aanleiding om items te verwijderen of buiten beschouwing te laten in de scoring. Leerlingen die van elkaar verschillen in thuistaal, leerlinggewicht en geslacht worden in de Eindtoets niet op oneigenlijke wijze benadeeld door de opgaven, dat wil zeggen door itemkenmerken die met de te meten vaardigheid zelf niets van doen hebben.

6.3 Criteriumvaliditeit

Criteriumvaliditeit heeft in eerste instantie betrekking op de relatie tussen test- of toetsscores en gedrag buiten de concrete testsituatie. Of, zoals het in het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer, & Sijtsma, 2010) wordt gesteld "in hoeverre de testscore een goede voorspeller is van niet-testgedrag (retrospectief, gelijktijdig of predictief)" (p. 43). In de Eindtoets Basisonderwijs worden leervorderingen meetbaar gemaakt, dus bij "niet-testgedrag" moet men in eerste instantie denken aan de prestaties die de leerling op school laat zien. Het zou getuigen van criteriumvaliditeit wanneer de score op de Eindtoets een sterk verband zou laten zien met indicatoren van eerdere schoolprestaties van de leerlingen ("retrospectief"). Te denken valt aan de scores op de Entreetoets. Deze relaties zijn in paragraaf 6.2 al aan de orde geweest en de conclusie was daar dat de samenhang van de Eindtoetsscores met eerdere metingen van schoolprestaties zeer hoog was. Verder mag men verwachten dat de score op de Eindtoets een sterke samenhang vertoont met het onafhankelijk advies dat de leerkracht zelf uitbrengt met betrekking tot het meest wenselijke school- en brugklatype voor een leerling. Enerzijds werpt dit advies licht op de leervorderingen van de leerling. De leerkracht zal zich immers bij dit advies in belangrijke mate baseren op de schoolprestaties die de leerling (nu en in het verleden) levert en heeft geleverd. Anderzijds richt de leerkracht zijn blik hier op de toekomst: hoe schat ik de toekomstmogelijkheden van deze leerling in. Hij raakt daarmee uiteraard aan een van de belangrijkste functies van de Eindtoets. De Eindtoets rapporteert in het Leerlingrapport (zie *Handleiding Eindtoets Basisonderwijs 2010* en de *Interpretatie van het leerlingrapport*¹²) immers de scores¹³ en geeft aan wat de betekenis daarvan kan zijn voor de keuze van een school- of brugklatype. Met het oog hierop is het antwoord op de vraag wat deze gegevens waard zijn essentieel. Met andere woorden, hoe staat het met de predictieve validiteit van de Eindtoets? In overeenstemming met het aangegeven gebruiksdoel (het geven van onafhankelijke informatie met het oog op het best passende brugklatype in het voortgezet onderwijs) concentreren we ons vooral op de plaatsing in de brugklas en de doorstroom na een en twee jaar verblijf in het voortgezet onderwijs als criterium.

In het onderstaande zullen we de predictieve validiteit van de Eindtoets zo breed mogelijk onderbouwen.

¹² In 2003 nog uitgegeven als brochure, daarna beschikbaar gesteld via een speciaal voor scholen toegankelijke website.

¹³ In het rapport staat ook het onderdeel Wereldoriëntatie. We wijzen er nogmaals op dat we in deze verantwoording op het onderdeel Wereldoriëntatie niet ingaan omdat dit onderdeel niet in de predictie betrokken is.

In paragraaf 6.3.1 gaan we in op de relatie tussen het resultaat op de Eindtoets (de standaardscore en het daarop gebaseerde advies) en het door de leerkracht uitgebrachte advies voordat de toets werd afgenomen, c.q. de uitslag ervan bekend was. Daarbij zullen we ook de relatie met intelligentie leggen (in de vorm van scores op de Cito Intelligentietest EB); dit doen we in paragraaf 6.3.2. Ten slotte leggen we in paragraaf 6.3.3 de relatie tussen de standaardscore en de resultaten van het Toelatings- en doorstroomonderzoek (T&DO). Zo'n toelatings- en doorstroomonderzoek werd tot enkele jaren geleden eens in de twee tot drie jaar gehouden. Momenteel vindt het jaarlijks plaats. In deze verantwoording rapporteren we eerst het onderzoek uit 2002. Daarbij vatten we de resultaten samen zoals deze eerder zijn gepubliceerd in Staphorsius et al. (2007). Ook gaan we kort in op het onderzoek in 2005, dat via een iets andere opzet werd uitgevoerd. Vervolgens bespreken we de resultaten van Toelatings- en doorstroomonderzoek zoals gerapporteerd in de *Interpretatie van het leerlingrapport* in 2009. Ten slotte vermelden we hier de niet eerder gepubliceerde resultaten van het T&DO voor de EB-versie van 2010. Omdat het T&DO in feite een follow up studie vormt over meerdere jaren, is het niet mogelijk om alle resultaten te presenteren voor de toetsversie van 2010. Vanwege de grote equivalentie van de verschillende toetsversies is dat ook niet nodig: de resultaten met betrekking tot eerdere versies kunnen van toepassing worden geacht op de toets van 2010.

6.3.1 Relatie met het doorstroomadvies van de leerkracht

In 2010 vroegen we *voordat* het resultaat van de leerlingen op de EB 2010 bekend was aan de leerkrachten het schooladvies voor hun leerlingen aan Cito door te geven. Leerkrachten konden hun advies uitbrengen door keuze voor een of meerdere brugklastypen van voortgezet onderwijs (het zogenoemde doorstroomadvies). De brugklastypen zijn weergegeven in de kopjes van de kolommen van tabel 6.8. We bekijken de overeenkomst met vijf grovere categorieën die gehanteerd worden als basis voor het uitbrengen van het Cito-advies. Deze hebben betrekking op schooltypen. Deze categorieën en de daarmee overeenkomende intervallen van de standaardscores zijn weergegeven in de linkerkolom van tabel 6.8. De cellen in de tabel die we beschouwen als een overeenkomst tussen doorstroomadvies en het advies gebaseerd op de EB zijn grijs gearceerd. De getallen in de tabel zijn absolute frequenties¹⁴.

Voor 28.8% van de leerlingen was geen of geen eenduidig advies voorhanden (categorie '?'). Het betreft gevallen waarin geen of meer dan één advies is uitgebracht. In tabel 6.9 wordt aangegeven in hoeveel procent van de gevallen het leerkrachtadvies overeenkomt met het advies op basis van de score op de Eindtoets Basisonderwijs (de grijs gearceerde cellen in tabel 6.8), dan wel hoger of lager is. In de tabel zijn ter vergelijking soortgelijke gegevens opgenomen uit eerder onderzoek (Staphorsius et al., 2007) met betrekking tot de versies van 2005 en 2006.

In tabel 6.8 is duidelijk te zien dat de hoge frequenties van niet-overeenkomst zich groeperen in de cellen langs de grijs gearceerde diagonaal. Extremere gevallen van niet-overeenkomst komen dus relatief weinig voor. In ongeveer 70% van de gevallen is er sprake van overeenkomst tussen het doorstroomadvies en het advies dat wordt uitgebracht op basis van de standaardscore op de Eindtoets. In ongeveer 97% van de gevallen (96.7% in 2010) liggen beide adviezen niet verder dan één categorie uit elkaar. Een opvallende vaststelling is verder dat in de gevallen dat het advies van de leerkracht en het advies op basis van de score op de Eindtoets niet overeenkomen het doorstroomadvies van de leerkracht vaker hoger dan lager uitvalt. Wat hier de precieze reden voor is kan op grond van de tot nu toe beschikbare gegevens niet worden achterhaald, maar de cijfers lijken een vrij stabiele situatie te weerspiegelen (zie tabel 6.9).

¹⁴ Aantal totaal is exclusief leerlingen die in de zogenaamde naverwerking (vanwege bijvoorbeeld problemen met het inlezen van optisch leesbare formulieren) betrokken zijn.

Tabel 6.8 Overeenkomst van doorstroomadvies leerkracht en Cito-advies op basis van de Eindtoets 2010

	?	bb	bb/kb	kb	gt	gt/h	gt/h/v	havo	h/v	vwo	totaal
501-523 (bb)	4195	3552	2295	3071	1706	234	21	35	7	4	15120
524-529 (kb)	5237	800	1372	3429	5281	1342	81	281	52	4	17879
530-536 (gt)	9532	205	597	2262	9489	6153	460	3075	928	122	32823
537-544 (havo)	13144	23	62	305	3344	6098	828	9671	8529	3197	45301
545-550 (vwo)	7486	1	3	6	100	320	135	1711	5384	11032	26178
Totaal	39594	4581	4329	9073	20020	14147	1525	14773	14900	14359	137301

Tabel 6.9 Overeenkomst doorstroomadvies van de leerkracht en het op de Standaardscore gebaseerde Cito-advies

	2005	2006	2010
schooladvies hoger	19.54	17.43	19.90
komt overeen	70.20	71.40	69.89
schooladvies lager	10.25	11.17	10.11

Door middel van CatReg (DTSS, Universiteit Leiden), een programma dat het mogelijk maakt multiple regressieanalyse toe te passen op categorische data, is nagegaan in hoeverre het doorstroomadvies te voorspellen is op basis van de Eindtoets. Daarbij is de categorie '?' buiten beschouwing gelaten. De correlatie tussen de standaardscore op de Eindtoets en het doorstroomadvies van de leerkracht bedraagt .843. Op deze samenhang komen we in de volgende paragrafen nog terug.

6.3.2 Het doorstroomadvies: intelligentie en leervorderingen vergeleken

Leervorderingen en intelligentie

Alvorens in te gaan op de vraag hoe de relatie tussen de score op de Eindtoets en het doorstroomadvies van de leerkracht zich verhoudt tot de relatie tussen intelligentie en dit doorstroomadvies, gaan we eerst in op de relatie tussen intelligentie en leervorderingen. In het al eerder aangehaalde onderzoek van Van Dijk en Tellegen dat de ontwikkeling van de Nederlandse Intelligentietest voor Onderwijsniveau (NIO) ondersteunde, zijn in aanvulling op de eerder in paragraaf 6.2 beschreven gegevens, ook hierover gegevens te vinden die we hebben overgenomen in tabel 6.10. Het betreft de correlaties tussen scores op de NIO (Verbaal, Symbolisch en Totaal) en de Eindtoets (en de onderdelen daarvan).

Tabel 6.10 Correlaties NIO-factoren met onderdelen EB

NIO	Eindtoets Basisonderwijs			
	Taal	Rekenen-Wiskunde	Informatie-verwerking	Totaal
Verbaal	.63	.55	.68	.68
Symbolisch	.53	.74	.55	.68
Totaal	.67	.74	.71	.78

De structuur van de correlatiematrix komt overeen met wat theoretisch verwacht mag worden (zie met name de vet gedrukte correlaties). De gegevens zijn vermoedelijk¹⁵ gebaseerd op de Eindtoets van de jaargang 2001. In die jaargang heette het onderdeel Studievaardigheden nog Informatieverwerking. Begrijpend lezen maakte daar destijds – anders dan nu – nog deel van uit, hetgeen de hoge correlatie tussen NIO-Verbaal en Informatieverwerking verklaart.

Andere belangrijke informatie over de samenhang tussen intelligentie en leervorderingen (zoals gemeten met de Eindtoets) is terug te vinden in de wetenschappelijke verantwoording van de in 2009 uitgebrachte Intelligentietest Eindtoets Basisonderwijs (Van Boxtel & Hemker, 2009). Daarin worden de uitkomsten van analyses besproken op een bestand van 4275 leerlingen bij wie deze intelligentietest in de periode januari – maart 2009 werd afgenomen naast de Eindtoets. Alle correlaties tussen de scores op diverse niveaus zijn samengebracht in tabel 6.11, waarbij alleen leerlingen zijn meegenomen die ook de opgaven voor Wereldoriëntatie hadden gemaakt (N=3446).

Tabel 6.11 Samenhangen tussen scores op de Intelligentietest EB en scores op de Eindtoets Basisonderwijs 2009 (alle Pearson-correlaties significant bij $p < .01$; N=3446)

<i>Eindtoets Basisonderwijs</i>	Intelligentietest: subtests						Somscores		
	SYN	TE	SW	GF	DR	SF	Verbaal	Ruimte	Algemeen
Taal (1)	.65	.61	.43	.28	.34	.37	.72	.44	.65
Rekenen-wiskunde (2)	.42	.42	.33	.45	.42	.40	.49	.55	.61
Studievaardigheden (3)	.57	.55	.41	.37	.42	.42	.64	.53	.68
Wereldoriëntatie (4)	.60	.54	.40	.38	.37	.37	.65	.49	.65
Totaalscore excl. Wereldoriëntatie	.61	.59	.44	.41	.43	.44	.69	.56	.72
Totaalscore incl. Wereldoriëntatie	.65	.61	.44	.43	.45	.45	.73	.58	.75
<i>Deeltaken EB</i>									
Schrijven (1)	.53	.51	.37	.26	.34	.34	.60	.41	.57
Spellen werkwoorden (1)	.32	.27	.21	.12	.14	.16	.34	.18	.29
Spellen niet-werkwoorden (1)	.36	.30	.22	.13	.14	.17	.38	.19	.32
Begrijpend lezen (1)	.58	.56	.39	.25	.32	.37	.65	.41	.60
Woordenschat (1)	.58	.54	.34	.24	.25	.28	.62	.34	.54
Getallen en bewerkingen (2)	.34	.35	.29	.36	.36	.33	.41	.46	.51
Verhoudingen, breuken, procenten (2)	.41	.41	.32	.41	.38	.38	.48	.51	.58
Metten, meetkunde, tijd, geld (2)	.39	.40	.31	.47	.42	.38	.46	.55	.60
Hanteren studieteksten (3)	.47	.44	.33	.22	.27	.32	.52	.36	.50
Hanteren informatiebronnen (3)	.46	.44	.31	.23	.29	.28	.51	.35	.49
Kaartlezen (3)	.39	.37	.29	.36	.36	.33	.44	.46	.52
Schema's, tabellen, grafieken (3)	.44	.45	.34	.34	.38	.38	.52	.49	.58
Aardrijkskunde (4)	.51	.43	.34	.35	.33	.32	.54	.44	.56
Geschiedenis (4)	.53	.48	.33	.30	.30	.30	.57	.39	.54
Natuuronderwijs (4)	.52	.51	.37	.34	.34	.35	.59	.45	.59

¹⁵ In Van Dijk en Tellegen vinden we geen expliciete verwijzing.

De algemene intelligentie-index hangt .72 samen met de standaardscore op de EB en als de scores op Wereldoriëntatie worden verdisconteerd .75. Algemene intelligentie hangt ongeveer even sterk samen met alle hoofdonderdelen van de EB: correlaties variëren tussen .61 en .68. Verbale intelligentie hangt het sterkst samen met de prestaties op Taal (.72), gevolgd door Wereldoriëntatie (.65) en Studievaardigheden (.64) en het minst sterk met Rekenen-Wiskunde (.49). De samenhangen tussen ruimtelijke intelligentie enerzijds en EB-scores anderzijds zijn zwakker dan die voor verbale intelligentie. De hoogste samenhang is die met Rekenen-Wiskunde (.55), de laagste met Taal (.44). Dit patroon is geheel volgens verwachting. Ook op het niveau van de deeltaken van de EB zijn ze terug te vinden. We beperken ons tot de opmerking dat Begrijpend lezen voor de taalonderdelen de hoogste samenhang laat zien met verbale intelligentie (.65).

Het doorstroomadvies voorspeld door Intelligentietest en Eindtoets

In het bestand dat de basis vormde voor de analyses waarvan de uitkomsten hierboven zijn beschreven waren ook gegevens opgenomen over het advies van de leerkracht met betrekking tot het best passende type vervolgonderwijs. Het gaat om soortgelijke gegevens als waarvan sprake was in paragraaf 6.3.1 (zie tabel 6.8). Het advies was uitgebracht voordat de uitkomsten van de Intelligentietest EB en de Eindtoets bekend waren bij de leerkracht. De vraag laat zich stellen, in hoeverre de scores op de Eindtoets beter dan intelligentiescores de doorstroomadviezen van leerkrachten met betrekking tot het best passende type vervolgonderwijs kunnen voorspellen. Het uitgangspunt is dat leervorderingen de beste voorspeller vormen voor het best passende type vervolgonderwijs en dat intelligentie alleen toegevoegde waarde kan hebben in individuele gevallen (zie paragraaf 2.4.1 en Van Boxtel & Hemker, 2009).

In tabel 6.12 zijn per adviescategorie de gemiddelde scores voor de Intelligentietest en de Eindtoets aangegeven. In de twee kolommen daarnaast is steeds aangegeven wat het relatieve verschil (in termen van het aantal standaarddeviaties) met de adviescategorie daaronder is.

Tabel 6.12 Het doorstroomadvies van de leerkracht: scores op Intelligentietest en EB 2009 voor negen brugklastype alsmede de onderlinge relatieve verschillen (N=2898)

Doorstroomadvies leerkracht	Gemiddelde score			Relatief verschil in SD	
	IT	EB	N	IT	EB
vwo	116,6	546,5	342	0,49	0,40
havo/vwo	109,7	542,7	459	0,33	0,34
havo	105,2	539,5	373	0,27	0,21
gemengd theoretische leerweg/havo/vwo	101,4	537,6	63	0,11	0,21
gemengd theoretische leerweg/havo	100,0	535,6	424	0,28	0,45
gemengd theoretische leerweg	96,1	531,3	620	0,31	0,52
kaderberoepsgerichte leerweg	91,8	526,5	309	0,31	0,49
kaderberoeps-/basisberoepsgerichte leerweg	87,5	521,9	176	0,26	0,22
basisberoepsgerichte leerweg	83,9	519,9	132		
totaal	100.8	534.9	2898		

IT = Intelligentietest: algemene intelligentie-index

EB = Eindtoets Basisonderwijs 2009: standaardscore

Zowel voor de algemene intelligentie-index als voor de standaardscore op de Eindtoets lopen de gemiddelden op naarmate het advies een hoger onderwijstype betreft. De relatieve verschillen tussen de gemiddelden per adviescategorie wijken voor beide instrumenten licht van elkaar af. Het lijkt erop dat de Intelligentietest wat grotere verschillen laat zien in de hogere adviescategorieën, terwijl de Eindtoets dat lijkt te doen voor de lagere adviescategorieën.

Door middel van CatReg (DTSS, Universiteit Leiden) is nagegaan in hoeverre de voorspelling van het doorstroomadvies op basis van de Intelligentietest en de Eindtoets van elkaar verschilt. In tabel 6.13 zijn de gegevens terug te vinden voor elk van beide scores afzonderlijk en in combinatie. Het is duidelijk en conform de uitgangspunten dat de Eindtoets de beste voorspeller is. De correlatie met het doorstroomadvies is .83, terwijl die correlatie voor de Intelligentietest afzonderlijk 'slechts' .66 bedraagt. Wanneer beide instrumenten in de regressieanalyse worden opgenomen blijkt dat ze ieder hun eigen significante bijdrage leveren aan de voorspelbaarheid van het doorstroomadvies. De verklaarde variantie die bij gebruik van de Eindtoets alleen ook al 69% is, neemt daardoor echter niet toe. Deze uitkomsten zijn als volgt te interpreteren. De Intelligentietest is op zichzelf een goede voorspeller van het doorstroomadvies ($r=.66$), maar de voorspelbaarheid van het doorstroomadvies wordt aanzienlijk verbeterd door ook de Eindtoets Basisonderwijs af te nemen. Dat de verklaarde variantie van beide instrumenten tezamen niet beter is dan die van de Eindtoets alleen, moet niet zo geduid worden als zou de Intelligentietest geen toegevoegde waarde hebben. De gebruiker heeft immers nu de beschikking over twee resultaten die bij bepaalde subgroepen verschil laten zien en daarom ook diagnostische betekenis hebben.

Tabel 6.13 Voorspelbaarheid van het doorstroomadvies op basis van Intelligentietest en Eindtoets (versie 2009) afzonderlijk en in combinatie (CatReg)

Voorspeller	B gestandaardiseerd	P	Correlaties		R ²
			0-order	partieel	
Regressieanalyse met alleen de Intelligentietest als voorspeller					
* IT: algemene intelligentie-index	.655	< .0001	.655	-	.43
Regressieanalyse met alleen de Eindtoets Basisonderwijs als voorspeller					
* Eindtoets Basisonderwijs 2009	.828	< .0001	.828	-	.69
Regressieanalyse met IT en EB beide als voorspellers					
* IT: algemene intelligentie-index	.120	< .0001	.655	.149	
* Eindtoets Basisonderwijs 2009	.742	< .0001	.828	.682	.69

Deze toegevoegde waarde van de Intelligentietest wordt geïllustreerd in tabel 6.14, waarin de uitkomsten van analyses worden gepresenteerd die kunnen worden beschouwd als aanvullend op de regressieanalyse.

Tabel 6.14 Doorstroomadvies leerkracht in relatie tot het verschil in resultaat op de Intelligentietest (IT) en de Eindtoets Basisonderwijs 2009

Verschil EB versus IT	EB<IT		IT=EB		EB>IT		Totaal	
	Abs.	%	Abs.	%	Abs.	%	Abs.	%
Doorstroomadvies								
vmbo	203	67.9	990	41.4	44	21.1	1237	42.7
gt	30	10.0	408	17.1	49	23.4	487	16.8
havo/vwo	66	22.1	992	41.5	116	55.5	1174	40.5
Totaal	299	100.0	2390	100.0	209	100.0	2898	100.0

Voor deze analyse werd de score op de EB 2009 op dezelfde schaal gebracht als die van de Intelligentietest (IT). Vervolgens werd de verschillscore bepaald en werden de leerlingen in drie groepen onderverdeeld: leerlingen met een duidelijk hogere EB- dan IT-score (verschil >15), leerlingen met een duidelijk lagere EB- dan IT-score (verschil >15) en leerlingen met (ongeveer) gelijke scores (de overige

leerlingen). Vervolgens werden deze groepen gerelateerd aan het doorstroomadvies van de leerkracht (voor de overzichtelijkheid gereduceerd tot een drietal categorieën). Verwacht werd dat 'onderpresteerders' relatief vaker naar de lagere schooltypen worden verwezen, terwijl 'overpresteerders' relatief vaker naar hogere schooltypen worden verwezen. De tabel laat duidelijk zien dat dit inderdaad het geval is; de betreffende samenhang is significant bij $p < .0001$. Kennelijk baseren leerkrachten zich in hun advisering vooral op de schoolprestaties die leerlingen daadwerkelijk laten zien (dat is ook wat zij dagelijks waarnemen) en in mindere mate op de capaciteiten van de leerlingen in termen van hun intelligentie. Het verstrekken van aanvullende informatie over de intelligentie lijkt dus met name zinvol voor de groep leerlingen die niet presteert in overeenstemming met zijn mogelijkheden.

Natuurlijk is het in het kader van deze verantwoording interessant om na te gaan of de Eindtoets van 2010 eenzelfde beeld laat zien. Bij 3918 leerlingen is in 2010 naast de Eindtoets ook de Cito Intelligentietest EB afgenomen. De correlatie van .72 tussen Eindtoets (standaardscore exclusief Wereldoriëntatie) en IT is exact gelijk aan die van 2009. Ook de intercorrelaties tussen enerzijds de toetsonderdelen Taal en Rekenen-Wiskunde en anderzijds de score voor verbale en ruimtelijke intelligentie laten hetzelfde patroon zien als in 2009 (zie tabel 6.11; de correlaties verschillen niet méér dan .02 van elkaar).

Tabel 6.15 Voorspelbaarheid van het doorstroomadvies op basis van Intelligentietest en Eindtoets 2010 afzonderlijk en in combinatie (CatReg)

Voorspeller	B gestandaardiseerd	P	Correlaties		R ²
			0-order	partieel	
Regressieanalyse met alleen de Intelligentietest als voorspeller					
* IT: algemene intelligentie-index	.663	< .0001	.663	-	.44
Regressieanalyse met alleen de Eindtoets Basisonderwijs als voorspeller					
* Eindtoets Basisonderwijs 2010	.823	< .0001	.823	-	.69
Regressieanalyse met IT en EB beide als voorspellers					
* IT: algemene intelligentie-index	.155	< .0001	.663	.155	
* Eindtoets Basisonderwijs 2010	.712	< .0001	.823	.666	.69

De regressieanalyses (zie tabel 6.15) laten voor deze 2010-data exact hetzelfde beeld zien als voor de 2009-data in tabel 6.13. De conclusie is dan ook gelijklopend: de correlatie van de Eindtoets met het doorstroomadvies is .82¹⁶, terwijl die correlatie voor de Intelligentietest afzonderlijk 'slechts' .66 bedraagt.

6.3.3 De score op de Eindtoets en de feitelijke plaatsing en doorstroom

Het Toelatings- en doorstroomonderzoek van 2002 en 2005

We rapporteren hier de opzet en resultaten van het onderzoek naar de toelatings- en doorstroomgegevens van leerlingen die in 2001 de Eindtoets hebben gemaakt en dus in 2002 het eerste cursusjaar van het voortgezet onderwijs afsloten.

De opzet van het onderzoek in 2002 was als volgt:

- Aan basisscholen is gevraagd aan te geven op welke school voor voortgezet onderwijs hun leerlingen die in 2001 de EB maakten, zijn toegelaten;

¹⁶ Dit getal wijkt enigszins af van de eerder aangegeven .83 en .84, maar is op andere (2009), respectievelijk een deel van de data (2010) berekend.

- Aan scholen voor voortgezet onderwijs is gevraagd aan het begin van het tweede jaar aan te geven tot welk brugklatype de leerlingen in 2001 zijn toegelaten en in welke type ze aan het begin van het tweede leerjaar in 2002 zijn geplaatst.

Via ons verzoek aan de basisscholen konden we in 2002/2003 van 104 137 leerlingen achterhalen op welke school voor voortgezet onderwijs zij waren toegelaten. Van de scholen voor voortgezet onderwijs kregen we vervolgens van 68 547 leerlingen de toelatings- en doorstroomgegevens. Van deze leerlingen wisten we welke standardscore ze op de EB 2001 hadden behaald, in welk brugklatype ze waren toegelaten en of ze al dan niet waren bevorderd naar het tweede jaar en of dat in hetzelfde of in een hoger of lager brugklatype was dan waarin ze waren toegelaten aan het begin van het eerste leerjaar voortgezet onderwijs. In de brochure *Interpretatie van het Leerlingrapport 2003* werden de resultaten van het onderzoek gepresenteerd; wij geven ze hier kort weer.

Inspectie van deze toelatingsgegevens leert dat de overeenstemming tussen 'toelating' en 'verwachting op basis van de standardscore' hoog is. In 79% van de gevallen zijn leerlingen overeenkomstig het advies op het leerlingrapport van de EB 2001 toegelaten. In 10% van de gevallen zijn leerlingen in een 'lager' en in 11% in een 'hoger' brugklatype terechtgekomen dan door de EB 2001 voorspeld.

De doorstroomgegevens op langere termijn worden steeds in de brochure *Interpretatie van het Leerlingrapport* verwerkt: ze geven de leerkracht aan wat de kans is dat een leerling met een bepaalde standardscore niet zal worden bevorderd naar het tweede leerjaar in hetzelfde school- of brugklatype waartoe deze is toegelaten. Het blijkt dat de plaatsing in het tweede leerjaar van het voortgezet onderwijs in 68% van de gevallen overeenkomt met de plaatsing op basis van de score op de EB.

In 2005 is een soortgelijk onderzoek uitgevoerd via een andere logistieke opzet. Scholen voor voortgezet onderwijs werden direct bevraagd over de plaatsing en doorstroom van leerlingen. Bovendien leverden zij de gegevens niet langer aan met behulp van optisch leesbare formulieren zoals in 2002 en eerder gebeurde. De gegevens werden nu geleverd door middel van exports uit de schooladministratie-programma's. We konden uiteindelijk beschikken over toelatings- en doorstroomgegevens van 54 958 leerlingen. Het onderzoek uit 2005 repliceert de resultaten van 2002, net zoals de gegevens in 2002 een replicatie waren van de gegevens die in 1999 verzameld werden. Een gedetailleerdere rapportage van deze onderzoeken is te vinden in Staphorsius et al. (2007).

Het meest recente Toelatings- en doorstroomonderzoek

Het meest recente Toelatings- en doorstroomonderzoek is uitgevoerd tussen 2005 en 2008. De uitkomsten daarvan zijn toegepast in de interpretatiebrochure bij het Leerlingrapport bij de Eindtoets 2010 waarin de toetsresultaten van individuele leerlingen worden 'vertaald' naar een advies omtrent het best passende brugklatype in het voortgezet onderwijs. Omdat deze gegevens een goed beeld geven van de predictieve validiteit zullen zij hier wat uitgebreider worden besproken. Meer gedetailleerde informatie is te vinden in de brochure *Interpretatie van het Leerlingrapport* die als separate bijlage is toegevoegd.

De resultaten van 135 442 van de 153 728 leerlingen die in 2005 aan de Eindtoets Basisonderwijs hebben deelgenomen (88,1%) zijn gekoppeld aan —bij het Centraal Bureau voor de Statistiek (CBS) beschikbare — individuele gegevens over het voortgezet onderwijs. Cito verkreeg op deze manier in het schooljaar 2005-2006 gegevens over de instroom van deze leerlingen in het eerste leerjaar van het voortgezet onderwijs. Daarvan werden 17 109 leerlingen (17,1%) niet meegenomen in verdere analyses omdat zij in het eerste leerjaar volgens het CBS in de brede brugklatypes 'overig vmbo' en 'vmbo/avo' of in het praktijkonderwijs werden geplaatst. In 2006-2007 en 2007-2008 werden daar respectievelijk de gegevens over de doorstroom naar het tweede en derde leerjaar aan toegevoegd. Ook konden eventuele doubleurs worden geïdentificeerd. Daarbij was het informatieverlies zeer gering: van slechts 0,6, respectievelijk 0,8% van de leerlingen in het bestand waren geen gegevens over het tweede, respectievelijk derde leerjaar bekend. Bij de interpretatie van de score op de Eindtoets kan onderscheid gemaakt worden naar schooltype en naar brugklatype (voor de score-intervallen zie tabel 4.5 en 4.6). Daarnaast verwijzen wij ook terug naar figuur 4.4 waarin grafisch is weergegeven hoe de prestatieverdeling voor de verschillende onderscheiden brugklatypes er bij de instroom in het eerste leerjaar in het schooljaar 2005-2006 uitzag. Uit de figuur is af te leiden dat er een sterke samenhang is tussen de score die behaald is op de Eindtoets en de plaatsing in

het eerste leerjaar. Dat is ook op een andere manier getalsmatig weer te geven. Van de totale groep leerlingen in het doorstroomonderzoek heeft 80% een plaatsing gekregen die in overeenstemming is met het geadviseerde schooltype. Bij 13% van de leerlingen adviseert de Eindtoets 'hoger', bij 7% 'lager'. Op deze 'afwijkende' plaatsingen komen we zo dadelijk nog terug. In tabel 6.16 is te zien hoe de overeenstemming tussen advies en feitelijke plaatsing er per schooltype uit ziet.

Tabel 6.16 *Percentage leerlingen zoals geplaatst in de onderscheiden schooltypen in relatie tot het advies op basis van de score op de Eindtoets 2005*

Categorie	Advies	Lager	Overeenkomstig	Hoger
501-523	bb	---	78	22
524-528	kb	6	44	50
529-536	gt	12	74	14
537-544	havo	6	87	7
545-550	vwo	2	98	---

Het is duidelijk dat de overeenkomst het geringst is bij de adviescategorie kaderberoepsgerichte leerweg. De overeenkomst met de feitelijke plaatsing is hier 44%, waarbij 50% van de leerlingen met dit advies in een hoger schooltype instroomt. Daarnaast is te zien dat de overeenkomst tussen plaatsing en advies groter wordt naarmate het onderwijsniveau hoger is; het percentage lager geplaatste leerlingen neemt daarbij af.

Tabel 6.17 geeft de relatie tussen enerzijds standardscore en advies en anderzijds de plaatsing in brugklustypen weer. De tabel is het best rijgewijs te interpreteren omdat de percentages van links naar rechts optellen tot honderd. Daarbij zijn de cellen waarin het advies (naar schooltype) en de plaatsing overeenstemmen grijs gearceerd. De tabel maakt duidelijk dat per adviescategorie de hoogste percentages steeds in het gearceerde gebied liggen (dat duidt op overeenstemming). Zoals al eerder aangegeven is de overeenstemming het geringst bij het advies kaderberoepsgerichte leerweg en is de overeenstemming het grootste bij adviezen havo en vwo. De percentages in de cellen buiten het gearceerde gebied zijn over het algemeen laag en naarmate de cellen verder van de diagonaal afliggen naderen de percentages tot nul.

Tabel 6.17 *Percentage leerlingen zoals geplaatst in de onderscheiden brugklustypen in relatie tot het advies op basis van de score op de Eindtoets (versie 2005)*

Categorie	Advies	bb	bb/kb	kb	gt	gt/h	gt/h/v	havo	h/v	vwo
501-523	bb	32	46	10	8	2	2	0	0	0
524-529	kb	6	28	17	26	13	9	0	1	0
530-536	gt	1	6	5	27	29	18	2	12	0
537-544	havo	0	1	0	5	14	18	4	51	7
545-550	vwo	0	0	0	0	0	13	1	42	43

De tabel maakt duidelijk dat er een grote samenhang tussen de score op de Eindtoets Basisonderwijs en de daarop gebaseerde adviezen enerzijds en de feitelijke plaatsing in het voortgezet onderwijs anderzijds.

De meest recente gegevens over plaatsing en doorstroom

De gegevens die in de vorige paragraaf werden besproken zijn de gegevens die werden gehanteerd om de *Interpretatie van het Leerlingrapport* bij de Eindtoets van 2010 te onderbouwen. Daarbij moet Cito dus teruggrijpen naar Eindtoetsversies van enkele jaren tevoren. Dat is onontkoombaar als men het advies naar aanleiding van de standardscore wil kunnen baseren op de feitelijke plaatsing en doorstroming van de leerlingen. Dit is geen enkel probleem omdat de versies over tijd volledig equivalent genoemd kunnen worden. Door met enige regelmaat deze TD&O's uit te voeren houdt Cito de vinger aan de maatschappelijke pols. Daarbij valt het op hoe gering de verschuivingen tussen de verschillende TD&O's zijn. Niettemin voert Cito de laatste jaren elk jaar een TD&O uit (en niet om de drie jaar zoals voorheen het

geval was). Bij het schrijven van deze wetenschappelijke verantwoording waren de nieuwste gegevens beschikbaar die zullen worden gebruikt bij de Interpretatie van het Leerlingrapport versie 2011. De percentages zijn weergegeven in de tabellen 6.18 en 6.19 die verder volledig analoog zijn aan de tabellen 6.16 en 6.17. Nadere bespreking van de tabellen kunnen we achterwege laten omdat de percentages nauwelijks verschillen van die van een jaar eerder. We kunnen er aan toevoegen dat de overeenstemming tussen het advies op basis van de score op de Eindtoets en de feitelijke plaatsing nog iets groter dan die tussen score op de Eindtoets en het doorstroomadvies van de leerkracht (zie paragraaf 6.3.1). Dit blijkt ook uit de regressie die analoog is uitgevoerd aan de analyse die is gerapporteerd in tabel 6.10, waarbij deze keer met behulp van CatReg niet het (onafhankelijk door de leerkracht uitgebrachte) doorstroomadvies maar de feitelijke plaatsing in het voortgezet onderwijs werd voorspeld op basis van de standaardscore op de Eindtoets. De correlatie blijkt nog iets hoger te zijn dan bij het leerkrachtadvies, namelijk .853.

Deze gegevens wettigen de veronderstelling dat de samenhang tussen de adviezen op basis van de Eindtoets 2010 en de feitelijke plaatsing van leerlingen in het voortgezet onderwijs eenzelfde beeld laat zien. Uitkomsten van analyses op basis van data die werden verzameld met equivalente eerdere versies van de Eindtoets kunnen op deze manier worden gebruikt ter ondersteuning van de predictieve validiteit van elke nieuwe (equivalente) versie.

Tabel 6.18 Percentage leerlingen zoals geplaatst in de onderscheiden schooltypen in relatie tot het advies op basis van de score op de Eindtoets 2006

Categorie	Advies	Lager	Overeenkomstig	Hoger
501-523	bb	---	76	24
524-528	kb	6	48	46
529-536	gt	12	75	13
537-544	havo	6	86	8
545-550	vwo	2	98	---

Tabel 6.19 Percentage leerlingen zoals geplaatst in de onderscheiden brugklastypen in relatie tot het advies op basis van de score op de Eindtoets 2006

Categorie	Advies	bb	bb/kb	kb	gt	gt/h	gt/h/v	havo	h/v	vwo
501-523	bb	32	44	12	7	2	2	0	0	0
524-529	kb	6	29	19	26	11	7	0	1	0
530-536	gt	1	6	5	28	31	16	3	11	0
537-544	havo	0	1	0	5	14	16	5	51	8
545-550	vwo	0	0	0	0	1	11	1	39	48

Het Cito-advies en niveauwisselingen

Bovenstaande analyses roepen de vraag op in hoeverre het afwijken van het Cito-advies bij de feitelijke toelating en plaatsing in het voortgezet onderwijs consequenties heeft voor de gang van de leerling door het onderwijs. Voor zover de verzamelde gegevens met betrekking tot leerjaar 2 en 3 in het hierboven uitgebreid besproken T&DO (2005-2008) dat toelieten, valt hier het nodige over te zeggen.

Het onderstaande is gebaseerd op resultaten zoals die gerapporteerd worden in Stroucken, Takkenberg & Béguin (2008). Aan deze publicatie zijn de volgende conclusies te ontleen:

- Jongens en meisjes wijken ongeveer even vaak af van het Cito-advies, maar meisjes gaan vaker dan jongens naar een brugklas van hoger niveau, terwijl dit voor jongens net omgekeerd is. Dat leidt er toe dat meisjes vaker havo of vwo volgen terwijl hun Cito standaardscore gemiddeld wat lager is dan die van jongens.
- Bij een Cito-advies havo of gemengde / theoretische leerweg gaan autochtone leerlingen iets vaker dan (niet-Westers) allochtone leerlingen naar een lager niveau in de brugklas (respectievelijk 8% tegenover 5%). Bij het Cito-advies kaderberoepsgerichte leerweg daarentegen gaan autochtone leerlingen wat vaker dan allochtone leerlingen naar een hoger brugklasniveau (27% tegen 24%).

- Leerlingen van wie de ouders een hoog inkomensniveau hebben, kiezen vaker voor een hoger brugklasniveau dan door Cito geadviseerd en minder vaak voor een lager brugklasniveau. Dat valt het meest op bij leerlingen met een advies kaderberoepsgerichte leerweg (37% uit de hoogste inkomensklasse naar een hoger brugklasniveau, tegen 22% uit de laagste inkomensklasse).
- Wisseling van brugklasniveau na het eerste leerjaar komt relatief weinig voor. Slechts 3% zit na het eerste jaar in een hoger niveau ('opstroom') en eveneens 3% in een lager niveau ('afstroom').
- Ook leerlingen die afwijken van het Cito-advies komen voor de overgrote meerderheid in het tweede leerjaar op hetzelfde brugklasniveau terecht. Ze wisselen echter opvallend vaker van brugklasniveau dan leerlingen die het Cito-advies hebben gevolgd. Van laatstgenoemde subgroep gaat 2% naar een hoger en 2% naar een lager niveau. Van de leerlingen echter die in een hoger niveau zijn ingestroomd dan hun Cito-advies aangaf, stroomt 8% af, terwijl van de leerlingen die juist in een lager niveau zijn ingestroomd dan de Cito-score adviseerde, 13% alsnog naar een hoger niveau opstroomde. Hoewel deze resultaten geen duidelijke conclusies toelaten met betrekking tot de factoren en beweegredenen die een rol spelen bij het afwijken van het Cito-advies, kan voorzichtig worden geconcludeerd dat afwijkingen bij de overgang van leerjaar 1 naar leerjaar 2 enigszins lijken te worden gecorrigeerd in de richting van het op basis van de Eindtoets uitgebrachte advies.

7 Samenvatting

Deze wetenschappelijke verantwoording betreft de Cito Eindtoets Basisonderwijs, meer specifiek de versie van 2010.

De primaire functie van de Eindtoets Basisonderwijs is beschreven als: het geven van onafhankelijke informatie die de keuze van een school- of brugklastype voor voortgezet onderwijs aan het einde van groep 8 in de basisschool kan ondersteunen. Dit gebeurt op basis van toetsing van de leervorderingen: datgene wat de leerling in vergelijking met andere leerlingen in acht jaar basisonderwijs heeft geleerd op het gebied van Taal, Rekenen-Wiskunde en Studievaardigheden. Het onderdeel Wereldoriëntatie is facultatief en wordt niet bij de primaire functie van de toets betrokken.

Alle leerlingen van groep 8 kunnen in principe aan de toets deelnemen, met uitzondering van (allochtone) leerlingen die vier jaar of korter in Nederland zijn en het Nederlands onvoldoende beheersen en leerlingen die naar verwachting naar het (voortgezet) speciaal onderwijs of naar het praktijkonderwijs zullen gaan. Voor de zwakke leerlingen zijn speciale en aanvullende versies beschikbaar waarvan de scores kunnen worden geëquivalereerd naar scores op de reguliere Eindtoets. Voor leerlingen met beperkingen zijn er speciale versies uitgebracht (ingesproken op cd of daisy-cd, vergrote versie, zwart-witversie, brailleversie). Ook is er een digitale versie beschikbaar die ook als inhaaltoets kan worden ingezet. Deze verantwoording heeft primair betrekking op de papieren versie.

In hoofdstuk 2 en 3 is de toets uitgebreid beschreven en inhoudelijk verantwoord. Daarbij is veel aandacht besteed aan de constructie-, kalibratie- en equivaleringsprocedures die ervoor zorgen dat de toets van jaar tot jaar equivalent is ten opzichte van zijn voorgangers. Dit ondanks het feit dat de inhoud van de toets in termen van concrete opgaven steeds volledig wordt vernieuwd. Ook is het meetmodel beschreven dat aan de toets en de onderdelen ervan ten grondslag ligt. De ingezette technische hulpmiddelen spelen een belangrijke rol, maar daarnaast is het van groot belang om de inhoud van de toets constant te houden. Dit gebeurt door de toetsinhoud minutieus voor te schrijven door middel van een gedetailleerde beschrijving van onderwijsdoelstellingen in het zogenoemde Doelenboek. Deze onderwijsdoelstellingen vormen een adequate afspiegeling van de communale onderwijsinhouden. Een samenvatting van het Doelenboek is opgenomen in hoofdstuk 3. Andere maatregelen die bijdragen aan de hoge kwaliteit en constantheid van de toetsinhoud zijn:

- Cito analyseert doelstellingen en methoden en gaat aan de hand van de resultaten na of de Eindtoets nog goed aansluit bij de inhoud van het onderwijs.
- De constructie van opgaven vindt plaats in commissies waarvan de leden ervaring hebben in het basisonderwijs.
- Er vindt proeftoetsing plaats van alle opgaven.
- Een onderdeel van de jaarlijkse proeftoets is het verzoek om commentaar van leerkrachten.

Het feit dat de inhoud van de Eindtoets op straat ligt en de constructie ervan als het ware plaatsvindt 'in een glazen huis' speelt ook een belangrijke rol. De laatste jaren staat de inhoud van de toets niet of nauwelijks ter discussie.

Alle maatregelen hebben tot gevolg dat de Eindtoets Basisonderwijs ieder jaar opnieuw een equivalente inhoudsvalide toets is, die qua inhoud en structuur een replicatie vormt van de voorafgaande versie. De toetsscores en de scores op de onderdelen ervan zijn licht linksscheef verdeeld omdat de moeilijkheidsgraad van de opgaven bewust wordt gekozen tussen .40 en .90 met gemiddelden van omstreeks .70. Er wordt gestreefd naar optimale homogeniteit; de gemiddelde itemtotaalcorrelaties liggen per toetsonderdeel tussen de .35 en .40. De onderdelen laten een vast patroon van intercorrelaties zien. De ruwe scores worden omgezet in percentielscores. De totaalscore wordt daarnaast omgezet in een standardscore met een range van 501 tot 550. Het gemiddelde ligt jaarlijks om en nabij de 535 met een standaarddeviatie tussen 9 en 10. De standardscore is niet opzet niet genormaliseerd.

Qua normering is de situatie waarin de Eindtoets zich bevindt speciaal te noemen. De toets wordt namelijk jaarlijks genormeerd op de totale groep van deelnemende leerlingen. In 2010 ging het daarbij om ruim 137 000 leerlingen (waarbij wordt afgezien van leerlingen die aan de diverse speciale versies van de toets hebben deelgenomen). Er is dus geen sprake van een steekproef. Wel is nagegaan in hoeverre de populatie van Eindtoetsgebruikers afwijkt van de totale populatie Nederlandse leerlingen in groep 8 van het regulier basisonderwijs. Zowel op school- als op leerlingniveau kon in hoofdstuk 4 worden aangetoond dat de populatie van Eindtoetsgebruikers een exacte afspiegeling is van de totale populatie op belangrijke (stratificatie)kenmerken als regio, sekse, leeftijd (leertijd), mate van verstedelijking, leerlinggewicht, schoolgrootte en percentage achterstandsleerlingen op school. Men zou daarom de groep van Eindtoetsgebruikers ook kunnen omschrijven als een (zeer omvangrijke) representatieve steekproef uit de populatie.

In hoofdstuk 4 is de kalibratie- en normeringsprocedure in stappen beschreven, waarbij aandacht is besteed aan de gehanteerde grafische en statistische hulpmiddelen. Er is uitgebreid aangetoond dat de kalibratie geslaagd is te noemen aan de hand van de grafische weergave van itemkarakteristieke curven, overschrijdingskansen van S-toetsen, R1c-waarden en de nauwkeurigheid van de schattingen van itemparameters op basis van 'c'. Ten slotte is aangegeven hoe de standardscore interpreteerbaar wordt gemaakt in termen van adviescategorieën (naar schooltype en brugklatype). Dit gebeurt door het inzetten van een jaarlijks Toelatings- en doorstroomonderzoek in het voortgezet onderwijs.

In hoofdstuk 5 is de betrouwbaarheid van de Eindtoets 2010 geëvalueerd. Dit gebeurde op klassieke wijze in de vorm van Cronbachs alfa. De toegepaste IRT-methoden gaven ons de mogelijkheid om een andere vorm van betrouwbaarheid te berekenen die in hoge mate vergelijkbaar is met Cronbachs alfa, namelijk MAcc. Ook werd een test-hertestonderzoek gesimuleerd met 150 000 'virtuele' leerlingen. Alle ingezette statistische procedures leidden tot vrijwel exact dezelfde waarde van .95 à .96 voor de totaalscore (standardscore). Dat is zeer hoog te noemen en ruimschoots voldoende om belangrijke selectie- en plaatsingsbeslissingen op te kunnen baseren. De betrouwbaarheid van het onderdeel Taal bedraagt .91, die van Rekenen-Wiskunde .91 à .92, die van Studievaardigheden .81 à .82 en die van Wereldoriëntatie .90 à .91. De met deze waarden overeenkomende 95%-betrouwbaarheidsintervallen zijn beschreven. Voorts gaven de ingezette IRT-methoden de mogelijkheid om inzicht te geven in de lokale betrouwbaarheid en nauwkeurigheid van de toets en de toetsonderdelen. De meetfouten werden grafisch weergegeven. De figuren maken duidelijk dat de meetfout het kleinst is in de lagere en gemiddelde vaardigheidsregionen. Dit gegeven vloeit voort uit de keuze van de constructeurs voor relatief gemakkelijke toetsen met een linksscheve verdeling. Het differentiërend vermogen en de nauwkeurigheid zijn daardoor in de lagere en gemiddelde regionen optimaal. De invloed van de meetfout op de indeling in adviescategorieën is aangegeven.

In hoofdstuk 6 is zeer uitgebreid ingegaan op zowel de begrips- als de criteriumvaliditeit.

Een basisvoorwaarde voor begripsvaliditeit is de geslaagde kalibratie van de toetsonderdelen.

Deze kalibratie zorgt ervoor dat de opgaven kunnen worden opgevat als representanten van unidimensionale vaardigheidsdimensies ('latente trekken'). Hierboven is de geslaagdheid van de kalibratie al besproken. De homogeniteit van de toetsopgaven in termen van r_{it} en Cronbachs alfa sluit aan bij de aangegeven uitstekende modelfit.

Ook de constructie- en equivaleringsprocedures dragen bij aan de begripsvaliditeit. Door de adequaatheid ervan zijn de verschillende versies van de Eindtoets op te vatten als equivalente 'soortgenoten'.

Deze adequaatheid werd nog eens gedemonstreerd aan de hand van de resultaten van de equivalering van 2009 en 2010 op basis van diverse stabiele steekproeven van scholen, de resultaten van de proeftoetsingen en het inzetten van ankeropgaven in de reguliere afname van de Eindtoets bij een steekproef van leerlingen.

Er is hierboven al gewezen op de stabiliteit in de structuur van de toets en de onderdelen daarvan.

De intercorrelaties laten van jaar tot jaar hetzelfde patroon zien. Daarbij valt op dat de correlaties voor de onderdelen van Taal wijzen op een verschil in semantische (Begrijpend lezen, Schrijven en Woordenschat) en niet-semantische onderdelen (Spelling). De onderdelen van Rekenen-Wiskunde laten traditioneel zeer hoge intercorrelaties zien van boven de .97 waardoor Rekenen-Wiskunde moet worden opgevat als een unidimensionale vaardigheid. De intercorrelaties voor Studievaardigheden zijn minder hoog, maar nog altijd

hoog. Alle correlaties leiden tot de conclusie dat aan de verschillende aspecten van leervorderingen (leerinhouden) een gemeenschappelijke leervorderingsfactor ten grondslag ligt met een zeer grote gedeelde variantie.

In drie verschillende onderzoeken kon de relatie tussen Eindtoets en intelligentie worden vastgesteld. Afhankelijk van de intelligentietest (NIO, Cito Intelligentietest Eindtoets Basisonderwijs) en de Eindtoetsversie (2001, 2009, 2010, met of zonder Wereldoriëntatie) lag de correlatie tussen .72 en .78. Voor wat betreft de toetsonderdelen Taal en Rekenen-Wiskunde enerzijds en de testscore voor verbale en ruimtelijk-symbolische intelligentie anderzijds werden verwachte patronen in de hoogte van de correlaties aangetroffen.

Op een groot bestand van ongeveer 107 000 leerlingen kon de Eindtoets 2010 worden gerelateerd aan de in groep 7 afgenomen Entreetoets. Deze toets kent voor een groot deel dezelfde opbouw als de Eindtoets. Ondanks de geheel verschillende samenstelling in opgaven en de tussenliggende periode van vijf à tien maanden is de correlatie van .87 tussen de totaalscores zeer hoog. Dit duidt enerzijds op een grote soortgenootvaliditeit, anderzijds op een grote stabiliteit in de (vaststelling van de) leervorderingen. Tevens wordt hiermee de adequaatheid van de constructie- en kalibratieprocedures en de kwaliteit van de Cito-itembanken onderstreept. Ook voor de toetsonderdelen Taal en Rekenen-Wiskunde is de soortgenootvaliditeit hoog met correlaties van .83 en .84. Er is sprake van divergente validiteit (voor zover daarvan sprake kan zijn gezien de grote gemeenschappelijke factor in leervorderingen). Het onderdeel Studievaardigheden verschilt te sterk qua opzet en inhoud om duidelijke conclusies toe te laten. Cito publiceert jaarlijks in de zogenoemde *Terugblik en resultaten* een overzicht van resultaten van analyses waarbij samenhangen met leerlingkenmerken, verschillen tussen groepen leerlingen en trends aan de orde worden gesteld. Een aantal van deze resultaten kan als ondersteunend voor de begripsvaliditeit worden opgevat omdat zij (theoretisch onderbouwde) verwachtingen ondersteunen. Tot de in dat opzicht interessantste uitkomsten behoren de volgende:

- Het gemiddelde van de vier grote steden (Amsterdam, Utrecht, Rotterdam en Den Haag), is net als voorgaande jaren ook in 2010 lager dan het landelijk gemiddelde (533.6 voor de G4 versus 535.8 landelijk). Dit wordt echter verklaard doordat de scholen in de grote steden relatief veel leerlingen met een hoger leerlinggewicht herbergen. En naarmate leerlingen een hoger leerlinggewicht hebben, daalt hun gemiddelde standaardscore op de Eindtoets aanzienlijk.
- Drie groepen leerlingen scoren duidelijk lager dan andere leerlingen. De eerste groep betreft leerlingen die aan het begin van groep 8 vier jaar of korter in Nederland zijn en de Nederlandse taal onvoldoende beheersen. De tweede groep bestaat uit leerlingen die naar verwachting naar het praktijkonderwijs of het (voortgezet) speciaal onderwijs gaan. De derde groep wordt gevormd door leerlingen die in aanmerking komen voor leerwegondersteunend onderwijs.
- Net als in voorgaande jaren behalen jongens gemiddeld iets hogere standaardscores dan meisjes. Dit geldt voor alle (leerling)gewichtscategorieën. Jongens en meisjes presteren op de verschillende onderdelen van de Eindtoets (gemiddeld genomen) duidelijk verschillend. Het verschil bij Taal is in het voordeel van de meisjes en is in 2010 toegenomen ten opzichte van 2009. Bij Rekenen-Wiskunde en Studievaardigheden is het verschil vrij stabiel in het voordeel van de jongens (op het onderdeel Studievaardigheden is het verschil overigens vrijwel nihil). Ook op het onderdeel Wereldoriëntatie, een facultatief onderdeel dat niet 'meetelt' voor de standaardscore, behalen jongens gemiddeld hogere scores dan meisjes.
- Op alle onderdelen scoort de groep vertraagde leerlingen gemiddeld duidelijk lager dan de groep leerlingen die normaal doorgestroomd is. Op alle onderdelen scoren de 'voorlijke' leerlingen hoger dan de overige leerlingen. Van de groep voorlijke leerlingen maken overigens meer meisjes dan jongens deel uit. Bij de vertraagde leerlingen is het net andersom: daar zijn weer meer jongens dan meisjes. Tot slot zijn, bij de groep leerlingen die normaal doorgestroomd is, de jongens gemiddeld ouder dan de meisjes.

De vraag is relevant of allochtone leerlingen, dan wel leerlingen met verschillende herkomst en thuistaal, op de Eindtoets anders scoren dan andere groepen leerlingen. Enerzijds is immers bekend dat allochtone leerlingen nog steeds een structurele achterstand laten zien ten opzichte van autochtone leerlingen, anderzijds is het maatschappelijk belang dat aan de Eindtoets Basisonderwijs wordt gehecht groot. Onderzoek naar enkele eerdere versies van de Eindtoets heeft laten zien dat er in beperkte mate sprake

was van Differential Item Functioning (DIF). Bij enkele opgaven was daarbij sprake van itembias (dat wil zeggen, de oorzaak van DIF was niet terug te voeren op het gemeten construct zelf). Deze beperkte itembias lijkt allochtone leerlingen echter noch te bevoordelen, noch te benadelen. Niettemin heeft Cito de afgelopen jaren een aantal opgaventypen aangepast en werd het onderzoek waarop de selectie van opgaven is gebaseerd zodanig uitgebreid dat het nog beter mogelijk is de selectie van opgaven met onbedoelde moeilijkheden te voorkomen.

Op de opgaven van de Eindtoets 2010 zijn DIF-analyses uitgevoerd voor de variabelen thuistaal, leerlinggewicht en sekse. Uit deze analyses is gebleken dat het aantal opgaven met DIF zeer beperkt is. Het aantal significante toetsingen is nauwelijks groter dan men op basis van toeval mag verwachten. In gevallen van significantie zijn de verschillen in functioneren niet goed interpreteerbaar en geeft ook de inhoud van de opgaven geen aanknopingspunten. De effecten op de toetsscore zijn bovendien verwaarloosbaar en niet systematisch, zodat kan worden geconcludeerd dat er geen specifieke categorieën leerlingen op een oneigenlijke manier worden benadeeld.

Met het oog op de primaire functie van de Eindtoets is de criteriumvaliditeit van de toets belangrijk. Criteriumvaliditeit heeft in eerste instantie betrekking op de relatie tussen test- of toetsscores en gedrag buiten de concrete testsituatie. Om zicht te krijgen op de criteriumvaliditeit zijn de scores op de Eindtoets gerelateerd aan het onafhankelijk uitgebrachte doorstroomadvies van de leerkracht. In ongeveer 70% van de gevallen is er sprake van overeenkomst tussen het doorstroomadvies en het advies dat wordt uitgebracht op basis van de standardscore op de Eindtoets. In ongeveer 97% van de gevallen (96.7% in 2010) liggen beide adviezen niet verder dan één categorie uit elkaar. Een opvallende vaststelling is verder dat in de gevallen dat het advies van de leerkracht en het advies op basis van de score op de Eindtoets niet overeenkomen het doorstroomadvies van de leerkracht vaker hoger dan lager uitvalt. De correlatie tussen beide adviezen bedraagt .84. Deze samenhang werd bevestigd door aanvullend onderzoek in 2009 en 2010 waarin ook de Cito Intelligentietest Eindtoets Basisonderwijs werd betrokken. De correlaties van .82 en .83 bleken aanzienlijk hoger dan de samenhang tussen intelligentie en doorstroomadvies. De Eindtoets heeft dus een aanmerkelijke meerwaarde bovenop het afnemen van een intelligentietest. Niettemin wordt een intelligentietest in aanvulling op de Eindtoets zinvol geacht, met name voor specifieke subgroepen leerlingen die niet in overeenstemming met hun mogelijkheden presteren op school. Uit deze onderzoeken blijkt ook dat de gemiddelde Eindtoetsscores oplopen voor groepen leerlingen naarmate hun doorstroomadvies een hoger school- en brugklastype betreft.

Met het oog op de predictieve validiteit ('over tijd') doet Cito regelmatig onderzoek naar de plaatsing en doorstroom van leerlingen in het voortgezet onderwijs. De overeenkomst tussen het advies op basis van de standardscore op de Cito Eindtoets en de feitelijke plaatsing in het voortgezet onderwijs is nog groter dan hierboven al is aangegeven voor het doorstroomadvies. In het schooljaar 2005-2006 heeft 80% van de leerlingen een plaatsing gekregen die in overeenstemming is met het geadviseerde schooltype. Bij 13% van de leerlingen adviseerde de Eindtoets 'hoger', bij 7% 'lager'. De overeenstemming is het geringst bij het advies kaderberoepsgerichte leerweg en het grootst bij de adviezen havo en vwo. Het meest recente Toelatings- en doorstroomonderzoek in 2006-2007 leverde vergelijkbare resultaten op.

Interessant is het gegeven dat in follow-uponderzoek blijkt dat ook leerlingen die afwijken van het Cito-advies voor de overgrote meerderheid in het tweede leerjaar op hetzelfde brugklasniveau terecht komen als in het eerste leerjaar. Ze wisselen echter opvallend vaker van brugklasniveau dan leerlingen die het Cito-advies hebben gevolgd. Van laatstgenoemde subgroep gaat 2% naar een hoger en 2% naar een lager niveau. Van de leerlingen echter die in een hoger niveau zijn ingestroomd dan hun Cito-advies aangaf, stroomt 8% af, terwijl van de leerlingen die juist in een lager niveau zijn ingestroomd dan de Cito-score adviseerde, 13% alsnog naar een hoger niveau opstroomde.

Er kan voorzichtig worden geconcludeerd dat afwijkingen bij de overgang van leerjaar 1 naar leerjaar 2 enigszins lijken te worden gecorrigeerd in de richting van het op basis van de Eindtoets uitgebrachte advies.

8 Literatuur

- Boland, Th., & Mommers, M.J.C. (1987). *Schoolvorderingentoetsen aan het eind van het basisonderwijs: verschillen en overeenkomsten*. Nijmegen: Katholieke Universiteit, Instituut voor Onderwijskunde.
- Boxtel, H.W. van, & Hemker, B.T. (2009). *Wetenschappelijke verantwoording van de Intelligentietest Eindtoets Basisonderwijs*. Arnhem: Cito.
- Cito (2010). *Terugblik en resultaten 2010*. Arnhem: Cito.
- Cito (2009). *Interpretatie van het Leerlingrapport*. Toelatings- en doorstroomgegevens van leerlingen die in 2005 aan de Eindtoets Basisonderwijs deelnamen. Arnhem: Cito.
- Deen, N. (1969). *Een halve eeuw onderwijsresearch in Nederland*. Groningen: Wolters Noordhoff.
- Dijk, H. van & Tellegen, P.J. (2004). *NIO Nederlandse Intelligentietest voor Onderwijsniveau*. Amsterdam: Boom test uitgevers.
- Evers, A., Braak, M.S.L., Frima, R.M., & Vliet-Mulder, J.C. van (2009). *COTAN Documentatie*. Amsterdam: Boom test uitgevers.
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests*. Geheel herziene versie met gewijzigde herdruk. Amsterdam: NIP/COTAN.
- Eggen, T.J.H.M. (1993). Itemresponstheorie en onvolledige gegevens. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 239-284). Arnhem: Cito.
- Engelen, R.J.H. en Eggen, T.J.H.M. (1993). Equivaleren. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 239-284). Arnhem: Cito.
- Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika* 53. pp. 525–546.
- Glas, C.A.W. & Verhelst, N.D., (1993). Een overzicht van itemresponsmodellen. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 179-238). Arnhem: Cito.
- Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item response Theory*. Newbury Park, CA: Sage.
- Heesters, K., Van Berkel, S., Van der Schoot, F. & Hemker, B. (2007). *Balans van het leesonderwijs aan het einde van de basisschool 4*. PPON-reeks nummer 33. Arnhem: Cito.
- Jansen, G.G.H., (1976). *Schaling en equivalering van de Eindtoets Basisonderwijs*. Arnhem: Cito.
- Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Ministerie van OCW. (1998). *Kerndoelen Basisonderwijs*. Zoetermeer: Ministerie van OCW.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

- Samenwerkende Instituten, RITP en Nutsseminarium voor Pedagogiek (1967), *Amsterdamse Schooltoetsen*. Groningen: Wolters.
- Schilt-Mol, T.M.M.L. van (2007). *Differential Item Functioning en itembias in de Cito-Eindtoets Basisonderwijs*. Amsterdam: Aksant.
- Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid: de ontwikkeling van een domeingericht meetinstrument*. Enschede: Universiteit Twente.
- Staphorsius, G. (red.) (2003). *Doelenboek*. Arnhem: Cito.
- Staphorsius, G. (red.) (2009). *Doelenboek*. Inhoudsverantwoording van de Eindtoets Basisonderwijs vanaf 2003. Arnhem: Cito.
- Staphorsius, G., Béguin, A., Engelen, R., Kleintjes, F., & Verhelst, N. (2007). *Eindtoets Basisonderwijs. Wetenschappelijke verantwoording op basis van een analyse van de met de jaargangen 2003-2006 verzamelde gegevens*. Arnhem: Cito.
- Staphorsius, G., Krom, R.S.H., Kleintjes, F.G.M. & Verhelst, N. (2001). *Begrijpend Lezen. Verslag van het kalibratie-, validerings- en normeringsonderzoek (wetenschappelijke verantwoording)*. Arnhem: Cito.
- Staphorsius, G., Verhelst, N. & Kleintjes, F.G.M. (2001). *Didactische Toetsen Schrijfvaardigheid. Verslag van het kalibratie-, validerings- en normeringsonderzoek (wetenschappelijke verantwoording)*. Arnhem: Cito.
- Staphorsius, G., & Verhelst, N.D. (2001). *Achtergronden*. In: *Verantwoording Toelatingstoets LWOO en PROO. Volg- en adviessysteem, Eerste Fase VO*. Cito: Arnhem.
- Stroucken, L., Takkenberg, T., & Béguin, A. (2008). Citotoets en de overgang van basisonderwijs naar voortgezet onderwijs. *CBS Sociaaleconomische trends, 2^e kwartaal 2008*, pp. 7-16
- Uiterwijk, H. (1994). *De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen*. Arnhem: Cito [proefschrift KUB]
- Uiterwijk, J.H., & Theunissen, T.J.J.M. (1996). *Verantwoording Eindtoets Basisonderwijs 1992*. Onderzoeksrapporten basis- en speciaal onderwijs. Arnhem: Cito.
- Uiterwijk, H., & Vallen, T. (1997). Onderzoek naar bias voor allochtone leerlingen in de Cito-Eindtoets Basisonderwijs. *Pedagogische Studiën*, 74 (1), pp. 21-32.
- Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, 22 (2), pp. 211-234.
- Verhelst, N.D. (1992). *Het één parameter model (OPLM). Een theoretische inleiding en een handleiding bij het computerprogramma*. Arnhem: Cito.
- Verhelst, N.D. (1993). Itemresponstheorie. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 83-178). Arnhem: Cito.
- Verhelst, N.D. & Eggen T.J.M.M. (1989). *Psychometrische en statistische aspecten van Peilingsonderzoek* (PPON-rapport, nr. 4). Arnhem: Cito.

- Verhelst, N.D. & Kleintjes, F.G.M. (1993). Toepassingen van itemresponsetheorie. In: T.J.H.M. Eggen en P.F. Sanders (Red.). *Psychometrie in de praktijk*. Arnhem: Cito.
- Verhelst, N.D., & Glas, C.A.W. (1995) The one parameter logistic model. In: G.H. Fischer & I.W. Molenaar (Eds.). *Rasch models: Foundations, recent developments and applications* (pp. 215-239). New York: Springer.
- Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1995). *OPLM: One Parameter Logistic Model. Computer program and manual*. Arnhem: Cito.
- Verhelst, N.D. & Kleintjes, F.G.M. (1993). Toepassingen van itemresponsetheorie. In: T.J.H.M. Eggen en P.F. Sanders (Red.). *Psychometrie in de praktijk*. Arnhem: Cito.
- Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.H.J.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model*. Measurement and Research Department Reports 91-10. Arnhem: Cito.
- Verstralen, H.H.F.M. (1997). *OPTAL: Inverse OPLAT and item and test characteristics in populations*. Arnhem: Cito.
- Wesdorp, H., (1974). *Het meten van de productief-schriftelijke taalvaardigheid*. Purmerend: Muusses.
- Wijnstra, J.M., (1984). *Verantwoording Eindtoets Basisonderwijs 1981* (met een historisch overzicht over de periode 1966 – 1980). Arnhem: Cito.

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
F (026) 352 11 35
klantenservice@cito.nl

Fotografie: Ron Steemers