Wetenschappelijke verantwoording Woordenschat groep 7 en 8

Saskia van Berkel, Maartje Hilte, Inge Groenen, Ronald Engelen

zeker weten

Wetenschappelijke verantwoording

Woordenschat groep 7 en 8

Saskia van Berkel
Maartje Hilte
Inge Groenen
Ronald Engelen

Cito
Arnhem, 2013

Inhoud

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de LVS-toetsen Woordenschat die deel uitmaken van het Cito Volgsysteem primair onderwijs (LOVS), te weten:
– de papieren toets Woordenschat voor groep 8;
– de digitale toetsen Woordenschat voor groep 7 en 8.
De papieren toetsen voor groep 3 tot en met 7 zijn reeds beoordeeld door de COTAN, evenals de digitale toetsen voor groep 3 tot en met 6.

De toetsen Woordenschat voor groep 3 tot en met 8 zijn te beschouwen als de tweede generatie toetsen in het Cito Volgsysteem primair onderwijs (LOVS). Nieuw ten opzichte van de eerste generatie is dat er naast een papieren ook een digitale versie beschikbaar is voor groep 3 tot en met 8 (groep 8 verschijnt in september 2013). Beide versies bevatten grotendeels dezelfde opgaven en zijn psychometrisch gezien als gelijkwaardige toetsen te beschouwen. Scholen maken zelf de keuze welke versie ze afnemen. Dit is mogelijk omdat de resultaten van de leerlingen op de beide versies steeds naar een en dezelfde vaardigheidsschaal te herleiden zijn.

Deze verantwoording biedt tezamen met de inhoud van de toetspakketten Woordenschat voor groep 7 en 8 alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit van de betreffende meetinstrumenten. Het genoemde materiaal maakt een beoordeling van de toetsen Woordenschat mogelijk op de volgende aspecten:
– Uitgangspunten van de toetsconstructie;
– De kwaliteit van het toetsmateriaal;
– De kwaliteit van de handleiding;
– Normen;
– Betrouwbaarheid;
– Validiteit.
Het laatstgenoemde aspect betreft alleen begripsvaliditeit en géén criteriumvaliditeit. Omdat de toetsen van het Cito Volgsysteem primair onderwijs (LOVS) niet bedoeld zijn voor 'voorspellend gebruik' is criterium-validiteit niet van toepassing.
Het voorliggende document heeft met name betrekking op de uitgangspunten van de constructie (de hoofdstukken 2 en 3), de normen (hoofdstuk 4), de betrouwbaarheid en meetnauwkeurigheid (hoofdstuk 5) en de begripsvaliditeit (hoofdstuk 6) van de digitale toetsen Woordenschat voor groep 7 en 8 én van de papieren toets voor groep 8. De kwaliteit van het toetsmateriaal en de handleiding is te bepalen door kennis te nemen van de inhoud van de toetspakketten.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

Onder woordenschat verstaan we de verzameling labels waarover leerlingen beschikken voor het begrijpen en gebruiken van taal.
De toetsen in de toetspakketten Woordenschat zijn bedoeld om vast te stellen hoe de receptieve woordenschat van leerlingen zich over de jaren heen ontwikkelt en welke verschillen er tussen leerlingen bestaan in zowel de breedte als de diepte van hun woordenschat. Met behulp van de toetsen kan, met andere woorden, onderscheid gemaakt worden tussen leerlingen met een beperkte, oppervlakkige woordenschat en leerlingen met een omvangrijke, diepe woordenschat.
De strategieën om woordbetekenis uit de context af te leiden of om de betekenis van woorden te onthouden worden niet expliciet bevraagd. De leerlingen laten indirect zien of ze voldoende woorden herkennen en of ze de relaties die er tussen woorden bestaan in voldoende mate beheersen (zie verder paragraaf 2.4.1).

2.2 Doelgroep

De toetsen Woordenschat voor groep 7 en 8 zijn primair bestemd voor en genormeerd bij leerlingen in de groepen 7 en 8 in het Nederlandse basisonderwijs. Voor de toetsen Woordenschat groep 3 tot en met 7 zijn de populatieparameters zowel op midden leerjaar als op einde leerjaar bepaald. Voor de toets voor groep 8 zijn de populatieparameters op begin en midden leerjaar bepaald. De toetsen kunnen desgewenst ook op andere momenten in het schooljaar worden afgenomen, maar dat maakt het moeilijker om uitspraken te doen over het niveau van de leerling ten opzichte van andere leerlingen in Nederland.
De toetsen zijn ook geschikt voor leerlingen op speciale scholen voor basisonderwijs en voor speciale leerlingen in het regulier basisonderwijs. In de handleiding bij de toetsen zijn – met het oog op het gebruik in deze doelgroepen – extra aanwijzingen opgenomen. Er zijn echter geen aparte referentiegegevens verzameld: voor speciale leerlingen zijn dezelfde normen van toepassing als voor leerlingen in het regulier basisonderwijs. Daardoor zijn de prestaties van beide doelgroepen (speciaal versus regulier) op de toetsen Woordenschat goed vergelijkbaar.
Voor leerlingen die nog maar pas in Nederland verblijven, zijn de toetsen echter ongeschikt: leerlingen dienen minstens vier jaar onderwijs in Nederland gevolgd te hebben alvorens de toetsen Woordenschat bij hen kunnen worden afgenomen.

2.3 Gebruiksdoel en functie

De toetsen Woordenschat in het Cito Volgsysteem primair onderwijs (LOVS) hebben twee doelen: niveaubepaling en progressiebepaling. Tevens bieden de toetsen de mogelijkheid de door de leerling gemaakte fouten te analyseren met het oog op het aanbieden van gerichte remediëring. Deze 'signalering' staat geheel los van de niveau- en progressiebepaling en is in de kalibratie- en normeringsonderzoeken niet wetenschappelijk getoetst.

Niveaubepaling
De toetsafnamen in het kader van woordenschat geven de leerkracht informatie over het niveau van de woordenschat van de leerlingen, individueel en als groep. Iedere behaalde vaardigheidsscore kan daartoe normgericht geïnterpreteerd worden op basis van de vaardigheidsverdeling in een adequate referentiegroep (zie paragraaf 4.2).

In de handleiding zijn twee niveau-indelingen opgenomen, waarmee de leerkracht de scores van een leerling kan vergelijken met die van een grote groep leerlingen.

De leerkracht kan een keuze maken uit:
– de indeling in de niveaus A tot en met E;
– de indeling in de niveaus I tot en met V.

Bij de indeling in de niveaus A tot en met E is de verdeling over de groepen als volgt:

Niveau	%	Interpretatie
A	**25**	De 25% hoogst scorende leerlingen
B	**25**	De 25% leerlingen die net boven tot ruim boven het landelijk gemiddelde scoren
C	**25**	De 25% leerlingen die net onder tot ruim onder het landelijk gemiddelde scoren
D	**15**	De 15% leerlingen die ruim onder het landelijk gemiddelde scoren
E	**10**	De 10% laagst scorende leerlingen

Bij de indeling in de niveaus I tot en met V wordt uitgegaan van vijf groepen van 20%:

Niveau	%	Interpretatie
I	**20**	De leerlingen die ver boven het gemiddelde scoren
II	**20**	De leerlingen die boven het gemiddelde scoren
III	**20**	De leerlingen die gemiddeld scoren
IV	**20**	De leerlingen die onder het gemiddelde scoren
V	**20**	De leerlingen die ver onder het gemiddelde scoren

In de eerste generatie toetsen van het leerlingvolgsysteem (LVS) werd uitsluitend de niveau-indeling A tot en met E gehanteerd. In de praktijk kent deze indeling echter een aantal nadelen.
De indeling is asymmetrisch opgebouwd. De niveaugroepen A, B en C bestrijken elk een kwart van de populatie en het vierde kwartiel is opgesplitst in twee subgroepen: D (15%) en E (10%). Bovendien interpreteert een groot aantal leerkrachten niveau C – het middelste niveau – als gemiddeld. Echter, de indeling A tot en met E toont geen gemiddelde groep leerlingen, maar alleen groepen die boven of onder het gemiddelde scoren.
Daarom is bij de tweede generatie toetsen van het leerlingvolgsysteem een indeling geïntroduceerd met de niveaus I tot en met V. Deze indeling is symmetrisch opgebouwd in vijf niveaugroepen van ieder 20%.
Dit heeft als voordeel dat er een ‘werkelijk’ middelste niveau onderscheiden wordt, niveaugroep III.
In strikt statistische zin kan echter ook bij niveaugroep III niet over *het gemiddelde niveau* worden gesproken. Het is theoretisch immers mogelijk dat bij een scheve verdeling de gemiddelde ruwe score niet eens in een dergelijke (middelste) groep ligt.

Progressiebepaling
De toetsen Woordenschat in het Cito Volgsysteem geven de leerkracht informatie over de ontwikkeling van de woordenschat van de leerlingen, individueel en als groep, gedurende (vrijwel) de gehele basisschool-periode. De toetsen geven antwoord op vragen als: is er sprake van vooruitgang, achteruitgang of van stabilisering? Is de vooruitgang – gelet op de gemiddelde vooruitgang in de populatie – volgens verwachting?

Het gehanteerde meetmodel (zie paragraaf 2.4.2) maakt het mogelijk om de scores van een leerling op verschillende toetsen, op verschillende momenten afgenomen, onderling te vergelijken. De ruwe scores op de toetsen – op basis van het aantal opgaven goed – zijn daartoe te transformeren in scores op één vaardigheidsschaal. Deze unidimensionele vaardigheidsschaal die aan de toetsen Woordenschat ten grondslag ligt, is ontwikkeld met behulp van het *One Parameter Logistic Model* (Verhelst, 1993; Verhelst & Glas, 1995; Verhelst, Glas & Verstralen, 1994).

'Signalering' via categorieënanalyse
Met behulp van de analyseformulieren bij de toetsen Woordenschat kan de leerkracht op eenvoudige wijze achterhalen met welke categorieën leerlingen problemen hebben. In de toetsen Woordenschat zijn twee categorieën onderscheiden, de categorie 'Betekenis' en de categorie 'Betekenisrelaties'. Leerlingen die één of beide categorieën onvoldoende beheersen, kunnen baat hebben bij extra instructie en gerichte oefeningen.
Er is geen kwalitatief of kwantitatief onderzoek gedaan naar het adequaat functioneren van de categorieën-analyse. De signalering via deze analyse heeft dan ook geen enkele wetenschappelijke status of pretentie. Haar enige functie is een handreiking bieden aan leerkrachten die gericht extra ondersteuning willen geven aan leerlingen bij wie de woordenschatontwikkeling achterblijft.

2.4 Theoretische inkadering

2.4.1 Theoretische inkadering: inhoudelijk

Wat is woordenschat?
Onder woordenschat verstaan we de verzameling labels waarover taalgebruikers beschikken voor het begrijpen en gebruiken van taal. Labels verwijzen naar concepten in het geheugen die samen iemands kennis van de wereld vormen. Een concept is een geheel van betekenissen, associaties, ideeën en beelden dat aan een woord verbonden is.
Elk woord heeft een woordvorm en een woordbetekenis. De woordvorm of het label is waarneembaar in tegenstelling tot woordbetekenissen, die zijn opgeslagen in het hoofd van de taalgebruiker. Met 'woord' of 'woorden' worden niet alleen 'losse' woorden, maar ook woordgroepen en woorden die samen een vaste verbinding vormen verstaan.

Woorden en hun betekenis(sen) worden voornamelijk op basis van andere woorden en betekenissen verworven. Ze zijn ingebed in het mentale lexicon. In dit begrippennetwerk zijn woorden knooppunten die relaties met andere knooppunten aangaan. Elk knooppunt bevat kennis over de klank, de betekenis, de grammaticale eigenschappen en de gebruiksmogelijkheden van woorden (Hilte, Van Berkel en Groenen, 2010; Verhallen en Verhallen, 1994).
Als de woordenschat van kinderen zich ontwikkelt, vindt een uitbreiding van het begrippennetwerk plaats. Er komen steeds nieuwe begrippen bij en er worden steeds meer relaties gelegd tussen al aanwezige begrippen. Het mentale lexicon van kinderen is echter nog niet zo gestructureerd en hiërarchisch opgebouwd als bij volwassenen. Zo associëren jonge kinderen *water* met *dorst hebben* en hebben oudere kinderen bij hetzelfde woord bijvoorbeeld de associatie met *milieu*. Dit heeft te maken met het gegeven dat het beheersen van woorden een cyclisch proces is dat zich niet in één keer ontwikkelt. Daarvoor is oefening en herhaalde toepassing in wisselende contexten nodig. Woorden en hun betekenis(sen) worden dus stapsgewijs begrepen, ingepast en toegepast, totdat ze geautomatiseerd zijn. Hoe vaker woorden gehoord worden, hoe duidelijker en preciezer hun betekenis kan worden vastgesteld. Bovendien wordt nieuwe kennis gemakkelijker verworven, omdat deze gekoppeld kan worden aan al bestaande kennis en aan bekende woorden. Voordat kinderen een beroep kunnen doen op abstracte begrippen (die uitsluitend op basis van taal worden opgebouwd), is het noodzakelijk dat de basiswoordenschat waarbij het vooral om concrete en alledaagse woorden gaat, voldoende ontwikkeld is. Hierin speelt vooral de breedte, maar ook de diepte van de woordenschat een belangrijke rol. Bij een brede woordenschat gaat het om het beheersen van veel verschillende woorden ofwel om de omvang van de woordenschat. Bij een diepe woordenschat

gaat het om de vraag: hoe goed ken je (betekenisaspecten van) woorden in relatie tot andere woorden? (zie o.a. Filipiak, 2006; Huizenga, 2005; Verhallen, 2006)
Vermeer (1997) beschrijft dat de breedte en de diepte van de woordenschat elkaar grotendeels overlappen. Kinderen die meer woorden kennen, kennen deze woorden vaak gedetailleerder en dieper. Dat komt omdat ze meer woorden tot hun beschikking hebben. In de bovenbouw van het basisonderwijs is een brede, oppervlakkige woordkennis niet toereikend en is diepe woordkennis noodzakelijk. Leerlingen in de hogere leerjaren moeten over een uitgebreid begrippennetwerk beschikken en over woordkennis die snel kan worden ingezet om verbanden en principes te begrijpen en problemen te kunnen oplossen.

Woordenschat in het onderwijs
Een van de doelstellingen van het onderwijs in de Nederlandse taal is dat leerlingen een adequate woordenschat verwerven. Woorden zijn immers de bouwstenen van de taal en liggen aan de basis van alledaagse en schoolse kennisoverdracht (zie onder meer Van den Nulft en Verhallen, 2002; Verhallen en Verhallen, 1994). Ze vervullen een centrale rol bij het verwerven en toegankelijk maken van kennis.
Alle leerstof is verpakt in woorden, leerkrachten geven woord voor woord uitleg, ze verwoorden verklaringen, brengen gedachteprocessen onder woorden en beschrijven verschijnselen en gebeurtenissen die zich elders in de ruimte en de tijd voordoen.
Hoewel kinderen een groot aantal woorden en woordbetekenissen al voor hun vierde levensjaar verwerven, is het vooral de school waar zij hun woordenschat vergroten; van ongeveer 3000 woorden in groep 1 tot ruim 25.000 woorden aan het einde van het voortgezet onderwijs (Huizenga, 2005; Schrooten en Vermeer, 1994).

Het beschikken over een brede en diepe woordenschat blijkt een belangrijke voorwaarde voor schoolsucces (Biemiller, 2010) en is daarmee van wezenlijk belang voor alle vakken die in het onderwijs aan bod komen. Zo biedt het voordelen bij het aanvankelijk lezen – als het te verklanken woord bekend is, maakt dat de directe woordherkenning gemakkelijker – maar ook bij het begrijpend lezen. Leerlingen met een beperkte woordenschat, slagen er minder goed in om kennis en vaardigheden op te doen. Omdat nuances in het schriftelijke en mondelinge taalaanbod hen veelal ontgaan, leren ze minder. In dit proces van cumulatieve achterstand spelen woorden, de betekenisdragers bij uitstek, een cruciale rol.
Schrooten en Vermeer (1994) geven aan dat de verschillen in woordenschat in groep 3 van het basisonderwijs nog weinig problemen lijken op te leveren. In groep 3 wordt namelijk relatief veel tijd besteed aan het aanvankelijk lezen, is het taalaanbod eenvoudig en lezen de leerlingen gemakkelijke en korte teksten. Deze bevatten veelal woorden waarvan leerlingen de betekenis al kennen. De zinnen zijn kort en er staan veel herhalingen in. Vanwege de nadruk op het technisch lezen hoeven leerlingen vaak niet eens te begrijpen wat ze lezen. Dit verandert als leerlingen in de midden- en bovenbouw meer en vlotter gaan lezen. De uitbreiding van de woordenschat wordt dan in toenemende mate bepaald door de geschreven taal. De variatie in leesteksten, onderwerpen en woorden neemt toe en leerlingen moeten het hoofd bieden aan weinig voorkomende, abstracte woorden en schrijftaalwoorden, aan moeilijker leesteksten met langere zinnen en complexere opdrachten. Leerlingen met een (boven)gemiddelde woordenschat versnellen, als ze de technische aspecten van het lezen achter de rug hebben, het tempo waarin ze hun woordenschat uitbreiden. Omdat ze al veel woorden en betekenissen kennen, kunnen ze nieuwe woorden en woordbetekenissen gemakkelijk inpassen bij wat ze al weten en kunnen ze al lezend de betekenis van onbekende woorden aan de hand van de context achterhalen. Op deze wijze leren ze nieuwe concepten en verbreden ze de betekenisnuances van woorden.
Dit staat in schril contrast tot leerlingen met een woordenschatachterstand. Voor deze leerlingen geldt dat teksten vaak zoveel onbekende woorden bevatten dat ze de betekenis niet uit de context kunnen afleiden (want ook daar staan onbekende woorden in). Deze leerlingen begrijpen daardoor nauwelijks waar teksten over gaan, nemen minder informatie tot zich, leren weinig of zelfs geen nieuwe woorden. De kans om achterop te raken is groot. Pas als ze op ongeveer tienjarige leeftijd een basiswoordenschat van zo'n 5.000 woorden hebben, versnelt hun woordgroei enigszins, hoewel in een langzamer tempo. Echter, de gevolgen van een geringe woordenschat zijn dan meestal al in de leerresultaten tot uitdrukking gekomen.
Er zijn voortdurend momenten waarop leerlingen in aanraking komen met nieuwe woorden. Als ze toevallig een nieuw woord tegenkomen en de noodzaak bestaat om achter de betekenis van dat woord te komen, is

er sprake van een incidentele woordleersituatie. Voor intentionele leersituaties geldt het omgekeerde: de leerkracht biedt dan, bijvoorbeeld in een woordenschatles, planmatig en systematisch nieuwe woorden aan. Het gaat daarbij meestal om het verbreden van de woordenschat, om het leren van steeds méér woorden. Op momenten dat leerlingen woorden in de les begrijpen of gebruiken, lijken deze vaak compleet verworven te zijn. Maar bij leerlingen met een woordenschatachterstand kan sprake zijn van onvoldoende of verkeerde betekenistoekenning. Zij kennen vaak maar één betekenis of alleen de letterlijke betekenis van een woord. Ook de betekenisaspecten die ze aan woorden toekennen zijn meestal minder abstract (Verhallen, 2006). Deze leerlingen hebben ook meer moeite om zich vaktermen en abstracte begrippen eigen te maken, waardoor ze de lessen minder gemakkelijk kunnen volgen en het hen moeite kost om nieuwe woorden te leren. Systematische aandacht voor woordenschatonderwijs is dan ook noodzakelijk. Dit zorgt ervoor dat leerlingen hun woordenschat vergroten en verdiepen, waardoor ze zich beter kunnen uitdrukken en beter uitgerust zijn om het onderwijs te volgen (Huizenga, 2005; Verhallen, 2006).

In het onderwijs moeten leerlingen de woorden die aan bod komen kunnen begrijpen én ze moeten zich kunnen uitdrukken. Als het begrijpen van taal centraal staat en leerlingen herkennen of interpreteren woorden, bijvoorbeeld bij het lezen van een verhaal of bij het luisteren naar instructies van de leerkracht, dan gaat het over de receptieve beheersing. Als leerlingen zelf iets in woorden uitdrukken, zoeken ze in hun mentale lexicon naar labels die het best uitdrukken wat ze mondeling of schriftelijk willen overbrengen en is er sprake van productie.
Het verschil tussen receptie en productie is minder groot dan vaak wordt aangenomen. Er is eerder sprake van een glijdende en overlappende schaal van receptief naar productief. Woordkennis is namelijk opgebouwd uit verschillende soorten kennis, met verschillende gradaties van receptieve en productieve beheersing, die niet gelijktijdig worden verworven. Zo kan het voorkomen dat een leerling wél de betekenis van een geïsoleerd woord kan geven, maar het niet in de juiste context kan gebruiken. Dikwijls gaat een productieve beheersing zelfs vooraf aan een echt goede receptieve beheersing (Filipiak, 2004).

2.4.2 Theoretische inkadering: psychometrisch

Opgavenbanken
Voor het samenstellen van toetsen voor het basisonderwijs beschikt Cito over opgavenbanken. Deze liggen onder meer ten grondslag aan de toetsen in het Cito Volgsysteem primair onderwijs, waaronder de LVS-toetsen, de Entreetoetsen en de Eindtoets Basisonderwijs. Voor de constructie van de toetsen Woordenschat hebben we gebruikgemaakt van de opgavenbank Woordenschat. Ook voor andere vakgebieden in het Cito Volgsysteem zoals Begrijpend lezen, Spelling, Rekenen-Wiskunde en Studie-vaardigheden zijn opgavenbanken in gebruik.

Een opgavenbank is nadrukkelijk niet 'zomaar' een verzameling opgaven waaruit een toetsconstructeur min of meer naar willekeur een aantal opgaven selecteert om een nieuwe toets samen te stellen. We geven hier kort aan wat de vereisten zijn om van een deugdelijke en psychometrisch goed gefundeerde opgavenbank te kunnen spreken.

Unidimensionaal continuüm
Het algemene uitgangspunt is dat de vaardigheid[1] in woordenschat kan 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 uit van de vaardigheid in woordenschat, waarbij een groter getal wijst op een grotere vaardigheid. Het doel van de meetprocedure – het afnemen van een 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 geeft aan hoe nauwkeurig die

[1] Het betreft hier een technische term die overeenkomt met een latente trek in de IRT.

schatting is, en heeft dus de status van een standaardfout, te vergelijken met de standaardmeetfout uit de klassieke testtheorie.

Latente vaardigheid
De antwoorden die een leerling geeft, worden beschouwd als indicatoren van de vaardigheid woordenschat, hetgeen ruwweg betekent dat men verwacht dat alle opgaven in de bank de woordenschat meten. De vaardigheid zelf wordt als niet observeerbaar beschouwd en daarom gewoonlijk omschreven als een latente vaardigheid.

'Moeilijkheid' in de Item Response Theorie
Hoewel opgaven dezelfde vaardigheid meten, kunnen ze toch systematisch van elkaar verschillen. Het belangrijkste verschil tussen de opgaven is hun moeilijkheidsgraad. In de klassieke testtheorie wordt moeilijkheidsgraad uitgedrukt met een zogenaamde p-waarde, de proportie correcte antwoorden op een opgave in een welbepaalde populatie van leerlingen. In de Item Response Theorie (IRT) die voor het construeren van de opgavenbanken werd gebruikt, hanteert men echter een andere definitie van moeilijkheid: ruwweg gesproken is het de mate van vaardigheid die nodig is om de opgave goed te kunnen beantwoorden. Dit verschil in definitie van de moeilijkheidsgraad tussen de klassieke theorie en IRT is uitermate belangrijk. Men kan verwachten dat de p-waarde van een opgave in groep 8 groter zal zijn dan in groep 7, waardoor duidelijk wordt dat de p-waarde een relatief begrip is: ze geeft de moeilijkheid aan van een opgave in een bepaalde populatie. Binnen de IRT is de moeilijkheid van een opgave 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 nodig om een opgave goed te kunnen beantwoorden) behoeft enige verdere uitwerking. Men zou deze omschrijving kunnen opvatten als een soort drempel. Heeft een leerling die mate van vaardigheid niet, dan kan hij de opgave niet 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 er uit volgt dat een leerling die een moeilijke opgave correct beantwoordt geen fout kan maken op een gemakkelijke opgave. Daarom wordt in de IRT een kansmodel gebruikt: hoe groter de vaardigheid, des te groter de kans dat een opgave juist wordt beantwoord. De moeilijkheidsgraad van een opgave wordt dan gedefinieerd als de mate van vaardigheid die nodig is om met een kans van precies een half een juist antwoord te kunnen geven.

Kalibratie
In het voorgaande zijn nogal wat veronderstellingen ingevoerd (unidimensionaliteit; alle opgaven zijn indicatoren voor dezelfde vaardigheid; kansmodel) die niet zonder meer voor waar kunnen worden aangenomen; we zullen methoden moeten bedenken om aan te tonen dat al die veronderstellingen deugdelijk zijn. Dit 'aantonen' gebeurt met statistische gereedschappen waarop we in deze paragraaf dieper zullen ingaan. Maar voor we de opgaven in een toets kunnen gebruiken moeten we ook proberen 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 die hiervoor wordt gebruikt, noemen we de kalibratiesteekproef.

Afnamedesigns
Meestal bevat een opgavenbank meer opgaven dan een doorsnee toets. Bij het uittesten van opgaven is het praktisch niet haalbaar, maar ook niet wenselijk om alle opgaven aan alle leerlingen voor te leggen. Elke leerling in de kalibratiesteekproef krijgt daarom slechts een gedeelte van de opgaven uit de opgaven-bank voorgelegd. Dit gedeeltelijk voorleggen gebeurt aan de hand van een 'onvolledig design' en 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).

Belangrijke implicaties gekalibreerde opgavenverzameling
Als we erin slagen de kalibratie met succes uit te voeren houden we een zogenaamde gekalibreerde opgavenbank over. In dat proces worden de opgaven die niet passen bij de verzameling verwijderd.
De opgavenbank bevat voor elke opgave niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen en de statistische zekerheid dat alle opgaven dezelfde vaardigheid aanspreken. Dit houdt onder meer het volgende in:

1 In principe kunnen we met een willekeurige selectie opgaven uit de opgavenbank de vaardigheid meten bij een willekeurige leerling. In principe, want een willekeurige toets die uit de opgavenbank wordt getrokken zal in de praktijk meestal niet voldoen omdat het meetresultaat (de schatting van de vaardigheid) onvoldoende nauwkeurig zal zijn. Willen we een nauwkeuriger meting (bij een gegeven aantal opgaven in de toets) dan zullen we de moeilijkheidsgraden van de opgaven in overeenstemming moeten brengen met het vaardigheidsniveau van de leerlingen.
   Het voorgaande geldt tevens voor de digitale opgaven. Ook deze zijn afkomstig uit de opgavenbank Woordenschat. Dus ook met een selectie van digitale opgaven kan de vaardigheid van een leerling bepaald worden. Al hetgeen dat geldt voor de 'papieren' opgaven uit de opgavenbank Woordenschat geldt ook voor de 'digitale' opgaven uit dezelfde opgavenbank.
2 We kunnen een schatting maken van de verdeling van de vaardigheid in een welomschreven populatie, door selecties van opgaven voor te leggen aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. In het geval van Cito Volgsysteem zijn dat steekproeven van leerlingen op de verschillende normeringsmomenten vanaf medio groep 3 (M3) tot medio groep 8 (M8). Daarbij maakt het, behoudens wat bij 1 is vermeld over nauwkeurigheid, niet uit welke selectie van opgaven bij een leerling binnen een normeringsgroep wordt afgenomen. Een van de eigenschappen van gekalibreerde opgavenbanken is immers dat met elke opgavenselectie de vaardigheid van leerlingen kan worden bepaald. Zie voor een voorbeeld hiervan Staphorsius (1994). In de praktijk komt dit meestal neer op het schatten van gemiddelde en standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met deze schattingen kunnen dan ook schattingen gemaakt worden van de percentielen in de populatie.
3 Aan leerlingen die niet behoren tot de betreffende referentiepopulatie kan dezelfde toets worden voorgelegd. De toetsscore wordt omgezet in een schatting van de vaardigheid en deze schatting kan geplaatst worden in de vaardigheidsverdeling van de populatie. Een leerling met achterstand in groep 7 kan een toets maken die normaliter aan leerlingen in groep 6 wordt voorgelegd. Zijn of haar vaardigheidsschatting kan behalve met de populatie van groep 7 ook vergeleken worden met de percentielen in de populatie van groep 6, in de vorm van bijvoorbeeld de uitspraak: "De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 6."
4 De vergelijking die in het voorgaande gemaakt is, kan evengoed plaatsvinden als de (achterstands)-leerling een andere toets (i.e. een selectie uit de opgavenbank) maakt dan de toets die normaliter aan de leerlingen in groep 6 wordt voorgelegd. Immers het kalibratieonderzoek heeft ons overtuigd dat alle opgaven dezelfde vaardigheid meten. Met een nieuwe toets meten we dus dezelfde vaardigheid, zodat schattingen die van verschillende toetsen afkomstig zijn zinvol met elkaar kunnen worden vergeleken.

Tot zover onze nadere bepaling van het begrip 'opgavenbank'. In de volgende hoofdstukken van deze verantwoording worden de begrippen die hierboven aan de orde zijn geweest nader uitgewerkt en toegelicht voor de opgavenbank Woordenschat. Voor de verantwoording van de constructie van deze opgavenbank verwijzen we naar hoofdstuk 3. In hoofdstuk 6 wordt de validering van de toets besproken.

Het gehanteerde meetmodel
In het normeringsonderzoek is gebruikgemaakt van een op de Item Respons Theorie (IRT) gebaseerd meetmodel zoals dat bij Cito gebruikelijk is. Dergelijke modellen verschillen in een aantal opzichten vrij sterk van de klassieke testtheorie (Verhelst, 1993; Verhelst en Glas, 1995). Bij de klassieke testtheorie staan de toets en de toetsscore centraal. Het theoretisch belangrijkste begrip in deze theorie is de

zogenaamde 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 weten we 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 opgave als een indicator voor de mate waarin die eigenschap aanwezig is. Het verband tussen eigenschap en antwoord op een opgave 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 kiezen we θ (theta). We wijzen erop dat θ niet rechtstreeks observeerbaar is. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd[2]. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 \mid \theta) \qquad (2.1)$$

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:

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

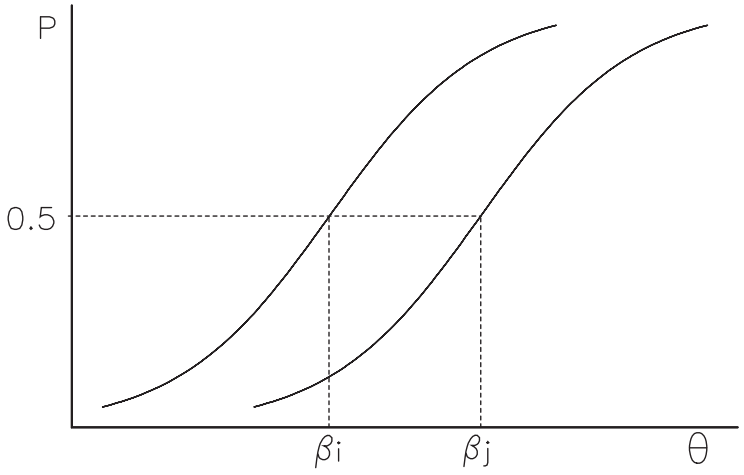
waarin β_i de moeilijkheidsparameter van item *i* is. Dat is een onbekende grootheid die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.1 voor twee opgaven, *i* en *j*, die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende 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, krijgen we:

$$f_i(\beta_i) = \frac{\exp(\beta_i - \beta_i)}{1 + \exp(\beta_i - \beta_i)} = \frac{1}{1 + 1} = \frac{1}{2} \qquad (3.3)$$

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*. We kunnen de parameter β_i dus terecht omschrijven als de moeilijkheidsparameter van item *i*. De implicatie van het bovenstaande is dat 'moeilijkheid' en 'vaardigheid' op dezelfde schaal liggen.

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

Figuur 2.1 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 2.1. Daaruit blijkt dat, voor welk vaardigheidsniveau dan ook, de kans om opgave *j* juist te beantwoorden steeds kleiner is dan de kans op een juist antwoord op opgave *i*. Daaruit volgt de statistisch te toetsen voorspelling dat de verwachte proportie juiste antwoorden op opgave *j* kleiner is dan op opgave *i* in een willekeurige steekproef van personen. Splitst men nu een grote steekproef in twee deelsteek-proeven, 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 blijkt echter lang niet altijd zo te zijn. Ook in ons geval niet. Veel van de opgaven blijken dan ook niet te kunnen worden beschreven met het Raschmodel. Daarom is bij dit instrument gekozen voor een ander IRT-model.

Alvorens het hier gebruikte model te introduceren, is 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 steekproefgrootheid ('sufficient statistic') bestaat voor de latente variabele θ, namelijk de ruwe score of het aantal correct beantwoorde opgaven. 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 opgave *i*, gegeven de ruwe score, een functie is die alleen afhankelijk is van de itemparameters en onafhankelijk van de waarde van θ[3]. 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).

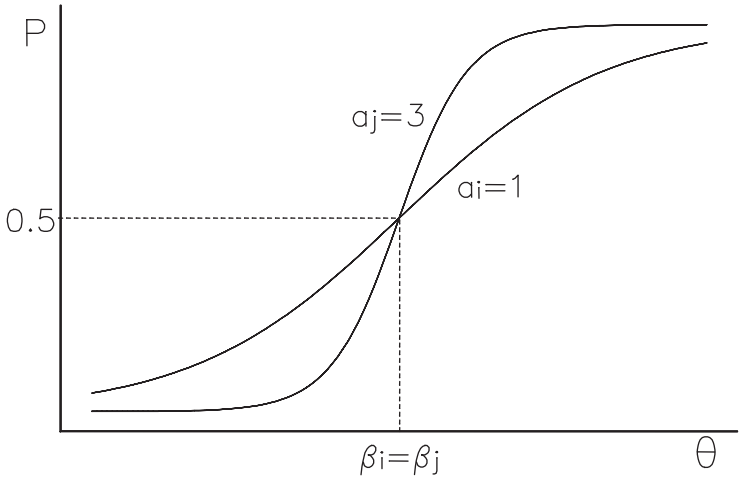
[3] Een gedetailleerde uiteenzetting hierover kan men vinden in Verhelst (1992).

De itemresponsfunctie van het OPLM is gegeven door:

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

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 2.2 is de itemresponscurve weergegeven van twee opgaven *i* en *j*, die even moeilijk zijn maar verschillend discrimineren.

Figuur 2.2 Twee itemresponscurven in het OPLM: zelfde moeilijkheid, verschillende discriminatie



De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas en Verstralen, 1995). Dit programma voert eveneens 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 opgaven volgens het OPLM is dan ook een iteratief proces waarin alternerend de modelfit van de opgaven 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 opgaven in het normeringsonderzoek zijn meerkeuzeopgaven, 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-schattingsmethode toe. De ongeschiktheid van het Raschmodel of OPLM voor meerkeuzevragen is echter relatief: indien de opgaven 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 meerkeuzeopgaven, 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 de populatieverdeling wordt gebruikgemaakt van de 'marginale grootste aannemelijkheidsmethode' (in het Engels: Marginal Maximum Likelihood, verder afgekort als MML). Deze schattingsmethode 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 een aselecte steekproef uit de populatie is. In hoofdstuk 4 tonen we aan dat aan deze voorwaarde voldaan is. Daardoor is het mogelijk om voor elk normeringsmoment een schatting te maken van deze (normaal verdeelde) vaardigheidsverdeling.

3 Beschrijving van de toets

3.1 Opbouw, afname, vorm en rapportage

Het toetspakket Woordenschat in het Cito Volgsysteem voor groep 7 bestaat uit twee toetsen: één toets die halverwege het schooljaar moet worden afgenomen (op het zogenaamde mediomoment) en één toets die aan het einde van het schooljaar moet worden afgenomen. Het toetspakket Woordenschat voor groep 8 bestaat daarentegen uit één toets die aan het begin óf halverwege het schooljaar moet worden afgenomen. In verband met de Eindtoets Basisonderwijs van Cito die jaarlijks in februari wordt afgenomen, was het vanuit praktisch oogpunt wenselijk om de toets Woordenschat groep 8 voorafgaand aan de afname van de Eindtoets te normeren. De normerings-/afnamemomenten van de toets Woordenschat groep 8 wijken om die reden af van de afnamemomenten van de toetsen voor groep 3 tot en met 7.
M.b.t. de groepen 7 en 8 gaat het dus om:
– de toetsen M7 (medio groep 7) en E7 (eind groep 7);
– de toets B8/M8 (begin of medio groep 8).
Van de papieren toetsen voor de groepen 7 en 8 is een digitale versie beschikbaar. De papieren en de digitale versie van de toetsen bevatten grotendeels dezelfde opgaven. Het aantal opgaven in de beide versies is gelijk.

Opbouw
De toetsen voor groep 7 bestaan per afnamemoment uit twee delen, gescheiden door een (korte) pauze. De beide delen moeten worden afgenomen. Elk deel bestaat uit 35 opgaven. De leerlingen maken dus 70 opgaven halverwege en 70 opgaven aan het einde van het schooljaar. De toets voor groep 8 bestaat eveneens uit 70 opgaven. Deze toets wordt aan het begin óf halverwege het schooljaar gemaakt.

Afname
De papieren toetsen worden klassikaal en schriftelijk gemaakt. De leerlingen krijgen een klassikale instructie en een aantal oefenopgaven, waarbij de verschillende opgaventypen die in de toets voorkomen besproken worden. De leerlingen lezen de opgaven (d.w.z. de vraag én de antwoordalternatieven) zelfstandig en noteren hun antwoord op een antwoordblad. De opgaven zijn afgestemd op het technisch leesniveau van de leerlingen in de desbetreffende groepen. Echter, voor leerlingen die moeite met lezen hebben, bestaat de mogelijkheid dat de leerkracht de vragen en bijbehorende antwoordalternatieven voorleest, de leerlingen lezen in dat geval mee.
De digitale toetsen worden individueel en aan de computer gemaakt. Afhankelijk van het aantal beschikbare computers kunnen meerdere leerlingen tegelijkertijd aan dezelfde toets werken. Eerst volgt een korte, algemene instructie die verband houdt met het bedienen van de computer. Vervolgens krijgen de leerlingen – voorafgaand aan elk opgaventype – een instructie over het opgaventype en een bijbehorende voorbeeldopgave. De eigenlijke opgaven worden getoond op het scherm, waarbij de mogelijkheid tot auditieve ondersteuning wordt geboden. Leerlingen die (een deel van) de opgaven willen beluisteren (bijvoorbeeld dyslectische leerlingen) hebben de mogelijkheid om op een ‘oortje’ te klikken. De betreffende opgave en de antwoordalternatieven worden dan voorgelezen.

Vorm
De toetsen voor groep 7 en 8 bestaan uit opgaven die de leerlingen zelfstandig moeten lezen.
De leerlingen lezen de gehele opgave, die bestaat uit een zin met een of meer stimuluswoorden en vier antwoordalternatieven. De stimuluswoorden maken deel uit van een vraag, van een zin of van een reeks woorden die moeten worden aangevuld. De antwoordalternatieven bestaan uit één of meerdere woorden of uit een korte zin.

Rapportage
De toetsen Woordenschat zijn zowel handmatig als via de computer te scoren en te analyseren. Voor het handmatig nakijken kunnen leerkrachten gebruikmaken van een lijst met goede antwoorden, die in de bijlage van de handleiding is opgenomen. Indien gewenst kan de leerkracht in het Computerprogramma LOVS de goede antwoorden aanklikken. Bij de digitale toetsen worden de antwoorden van de leerlingen door de computer gescoord. De leerkracht hoeft de toetsen dus niet na te kijken.
Na de toetsafname en de correctie van de leerlingantwoorden kunnen de toetsresultaten verwerkt worden op speciaal ontwikkelde rapportageformulieren. In de hoofdstukken 4 en 5 van de handleiding bij de toetspakketten Woordenschat en in de handleiding bij het Computerprogramma LOVS (zie de module Schoolzelfevaluatie) worden de mogelijkheden besproken om verschillende overzichten te maken, zoals leerlingrapporten, groepsrapporten, dwarsdoorsnedes en trendanalyses. Met behulp van deze overzichten kan de kwaliteit van het gegeven onderwijs ook op groeps- en schoolniveau geanalyseerd worden.

3.2 Inhoudsverantwoording

Allereerst komt in deze paragraaf het toetsen van de receptieve woordenschat aan bod. Aansluitend bespreken we de categorieën ‘Betekenis’ en ‘Betekenisrelaties’ en besteden we aandacht aan de inhouden en de opgaventypen zoals die in de toetsen aan bod komen. Ook gaan we in op de wijze waarop de selectie van de stimuluswoorden tot stand gekomen is. Tot slot komen de criteria die zijn gehanteerd bij het samenstellen van de toetsen Woordenschat aan de orde.
De informatie in deze paragraaf vormt een aanvulling op de inhoudsverantwoording die opgenomen is in de toetspakketten Woordenschat.

3.2.1 Woordenschat: een inhoudsanalyse

Het toetsen van de receptieve woordenschat
De toetsen Woordenschat doen een beroep op de receptieve woordenschat van de leerlingen. Omdat het kunnen begrijpen van mondelinge en schriftelijke informatie een uiterst belangrijke plaats in het onderwijs inneemt, ligt het voor de hand om de receptieve woordenschat van de leerlingen te meten. Dit houdt in dat de leerlingen zich niet in woorden hoeven uit te drukken, maar dat ze de woorden die ze krijgen aangeboden, moeten identificeren en herkennen om tot de betekenis van of betekenisrelaties tussen woorden te komen. Voor de toetsen Woordenschat betekent dit dat de leerlingen bij elke opgave een stimuluswoord en een aantal antwoorden krijgen aangereikt, waaruit ze een keuze moeten maken.
Het toetsen van de receptieve woordenschat heeft als belangrijk voordeel dat het nakijken van de toets en het bepalen van de toetsscores eenvoudig en objectief kan plaatsvinden.

De indeling ’Betekenis’ en ‘Betekenisrelaties’
Zoals beschreven in paragraaf 2.1 beogen de toetsen Woordenschat te meten of leerlingen de betekenis van woorden én de betekenisrelaties tussen woorden herkennen. Opgaven waarbij betekenistoekenning een centrale rol vervult, hebben we ondergebracht in de categorie ‘Betekenis’. Opgaven die betrekking hebben op relaties tussen woorden hebben we ingedeeld in de categorie ‘Betekenisrelaties’.
De indeling in ’Betekenis’ en ‘Betekenisrelaties’ sluit nauw aan bij de kwantitatieve en kwalitatieve aspecten van de woordenschat die in de literatuur worden beschreven als het gaat om de breedte en de diepte van de woordenschat (zie o.m. Filipiak, 2006; Huizenga, 2005; Verhallen, 2006). Deze aspecten zijn eveneens uitgewerkt in woordenschatoefeningen in verschillende taalmethoden voor het basisonderwijs. Voorbeelden hiervan zijn te vinden in methoden als Taalleesland, Taaljournaal, Taal op maat, Taalverhaal of Zin in Taal.

Tabel 3.1, 3.2 en 3.3 laten zien dat (ongeveer) de helft van alle opgaven in iedere toets behoort tot de categorie ‘Betekenis’ en de andere helft tot de categorie ‘Betekenisrelaties’. Dit komt nauw overeen met de opgavenverdeling zoals we die beoogd hadden, namelijk een fifty-fifty-verdeling. Bij de ontwikkeling van de woordenschat gaat het immers om de breedte én de diepte van de woordenschat. We vonden het daarom

niet meer dan logisch dat de beide aspecten een ongeveer even belangrijke plaats in de toetsen zouden moeten innemen.

Tabel 3.1 Aantal opgaven 'Betekenis' en 'Betekenisrelaties' in de papieren toets voor groep 8

Toets	Aantal opgaven 'Betekenis'	Aantal opgaven 'Betekenisrelaties'	Totaal aantal opgaven
B8/M8	35 (50%)	35 (50%)	70

Tabel 3.2 Aantal opgaven 'Betekenis' en 'Betekenisrelaties' in de digitale toetsen voor groep 7

Toets	Aantal opgaven 'Betekenis'	Aantal opgaven 'Betekenisrelaties'	Totaal aantal opgaven
M7	35 (50%)	35 (50%)	70
E7	32 (46%)	38 (54%)	70

Tabel 3.3 Aantal opgaven 'Betekenis' en 'Betekenisrelaties' in de digitale toets voor groep 8

Toets	Aantal opgaven 'Betekenis'	Aantal opgaven 'Betekenisrelaties'	Totaal aantal opgaven
B8/M8	38 (54%)	32 (46%)	70

Toetsinhouden en opgaventypen in de toetsen Woordenschat

Voorafgaand aan de opgavenconstructie voor de toetsen Woordenschat hebben we ons de volgende vragen gesteld:

– Welke inhouden willen we in de toetsen onderbrengen?
– Op welke aspecten van de woordenschat doen deze inhouden een beroep?
– Hoe verdelen we de verschillende inhouden en opgaventypen over de categorieën 'Betekenis' en 'Betekenisrelaties'?

Uitgaande van de eerste twee vragen zijn we tot de onderstaande indeling gekomen:

– Inhouden die refereren aan de kwantitatieve aspecten van de woordenschat: woorden labelen, woorden met eenzelfde betekenis, definities, beschrijvingen en belangrijke betekeniskenmerken;
– Inhouden die de kwalitatieve aspecten van de woordenschat representeren: tegenstellingen, betekenisveld, deel-geheelrelaties, gezamenlijke woordkenmerken en vergelijkingen.

Hoewel het psychometrisch gezien niet strikt noodzakelijk is om verschillende opgaventypen in de toetsen op te nemen, vonden we dat op basis van de inhoud wél van belang. De verschillende opgaventypen representeren immers de verschillende inhouden zoals we die in de literatuur en in taalmethoden hebben aangetroffen. Bovendien is een toets die uit meerdere opgaventypen bestaat voor de leerlingen aantrekkelijker en motiverender om te maken. Bij de verdeling in opgaventypen hebben we er daarom naar gestreefd een zo groot mogelijke spreiding in opgaventypen aan te brengen.

Wat de derde vraag betreft, kunnen we opmerken dat we geprobeerd hebben om de verschillende inhouden zo evenwichtig mogelijk te verdelen over de categorieën 'Betekenis' en 'Betekenisrelaties'.

Uit tabel 3.4, 3.5 en 3.6 is voor zowel de papieren als de digitale toetsen af te lezen welke inhouden en opgaventypen in de diverse toetsen zijn opgenomen. In de bijlagen 1a en 1b is van elk van de hier gepresenteerde opgaventypen een voorbeeld opgenomen.

Tabel 3.4 Toetsinhouden en opgaventypen m.b.t. de papieren toets voor groep 8

Toetsinhoud	**Opgaventype**	**B8/M8**
'Betekenis'		
Dezelfde betekenis	Wat is een ander woord voor …?	9
Definities	Wat is …?	3
Beschrijvingen	Wat betekent …?	9
	Welk woord past het best op de open plaats?	3
Belangrijke betekeniskenmerken	Waar gaat het bij / in de betekenis van … vooral om?	4
	Welke woorden zeggen het best iets over de betekenis van …?	7
	Totaal	*35*
'Betekenisrelaties'		
Tegenstellingen	Wat is het tegengestelde van …?	7
Betekenisveld	Wat past het best bij de betekenis van ...?	7
Gezamenlijke kenmerken (generalisatie)	Welk woord hoort er qua betekenis niet bij?	5
Gezamenlijke kenmerken (categorisatie)	Welk woord hoort niet bij de betekenis van ...?	4
Gezamenlijke kenmerken (betekeniscluster)	Vul het rijtje aan.	7
Trappen van vergelijking	Waar staan de woorden in de goede volgorde?	5
	Totaal	*35*

Tabel 3.5 Toetsinhouden en opgaventypen m.b.t. de digitale toetsen voor groep 7

Toetsinhoud	Opgaventype	M7	E7
'Betekenis'			
Dezelfde betekenis	Wat is een ander woord voor …?	8	9
Definities	Wat is …?	9	
Beschrijvingen	Wat betekent …?		8
	Een ... is iemand die ...	5	
	Met een … kun je …	5	
	Welk woord past het best op de open plaats?	8	
Belangrijke betekeniskenmerken	Waar gaat het bij … vooral om?		9
	Welke woorden zeggen het best iets over de betekenis van …?		6
	Totaal	*35*	*32*
'Betekenisrelaties'			
Tegenstellingen	Wat is het tegengestelde van …?	8	8
Betekenisveld	Wat past het best bij de betekenis van …?	6	3
Gezamenlijke kenmerken (generalisatie)	Welk woord hoort er qua betekenis niet bij …?	5	6
Gezamenlijke kenmerken (categorisatie)	Welk woord hoort niet bij de betekenis van …?	3	4
Gezamenlijke kenmerken (betekeniscluster)	Vul het rijtje aan.	5	10
Trappen van vergelijking	Waar staan de woorden in de goede volgorde?	8	7
	Totaal	*35*	*38*

Tabel 3.6 Toetsinhouden en opgaventypen m.b.t. de digitale toets voor groep 8

Toetsinhoud	Opgaventype	B8/M8
'Betekenis'		
Dezelfde betekenis	Wat is een ander woord voor …?	10
Definities	Wat is …?	3
Beschrijvingen	Wat betekent …?	9
	Welk woord past het best op de open plaats?	6
Belangrijke betekeniskenmerken	Waar gaat het in de betekenis van … vooral om?	4
	Welke woorden zeggen het best iets over de betekenis van …?	6
	Totaal	*38*
'Betekenisrelaties'		
Tegenstellingen	Wat is het tegengestelde van …?	8
Betekenisveld	Wat past het best bij de betekenis van ...?	7
Gezamenlijke kenmerken (generalisatie)	Welk woord hoort er qua betekenis niet bij?	4
Gezamenlijke kenmerken (categorisatie)	Welk woord hoort niet bij de betekenis van ...?	4
Gezamenlijke kenmerken (betekeniscluster)	Vul het rijtje aan.	5
Trappen van vergelijking	Waar staan de woorden in de goede volgorde?	4
	Totaal	*32*

3.2.2 De woorden in de toetsen Woordenschat

In het proces van opgavenconstructie stond voorop dat de toetsen (stimulus)woorden zouden moeten bevatten waar de leerlingen in groep 7 en 8 van het basisonderwijs daadwerkelijk mee in aanraking komen of zouden kunnen komen. De lijst 'Woorden in het basisonderwijs.15.000 woorden aangeboden aan leerlingen' van Schrooten en Vermeer (1994) bood hierbij uitkomst.
Deze descriptieve woordenlijst geeft een beeld van het mondelinge en schriftelijke taalaanbod op de basisschool aan vier- tot twaalfjarige leerlingen en bevat frequentie- en spreidingsgegevens bij 15.000 woorden, het minimale aantal woorden dat leerlingen gemiddeld receptief beheersen aan het einde van het basisonderwijs. De lijst van Schrooten en Vermeer is gebaseerd op het mondelinge taalaanbod van leerkrachten, alsmede op tekstmateriaal uit lesmethoden (taalmethoden Nederlands en zaakvakken-methoden) en jeugdboeken, waaronder prentenboeken en leesboeken. Gebruikers hebben de mogelijkheid om de woorden op de lijst te ordenen op jaargroep, frequentie en geometrisch gemiddelde (een maat waarin frequentie en spreiding verenigd zijn). Hoewel de lijst inmiddels al heel wat jaren in omloop is, was er ten tijde van de opgavenconstructie geen recenter en zeker ook geen gedetailleerder alternatief voor handen.

Het proces van woordselectie
Voor de opgavenconstructie voor de toetsen Woordenschat wilden we over een overzicht van stimulus-woorden beschikken met woorden passend bij het niveau van de leerlingen in de groepen 7 en 8.
Daartoe hebben we een aantal stappen doorlopen:

- Allereerst hebben we de eerste 3000 woorden op de lijst van Schrooten en Vermeer uitgesloten. Ervan uitgaande dat deze woorden – die zeer regelmatig in allerlei contexten en situaties voorkomen en gerekend kunnen worden tot de basiswoordenschat – bij alle leerlingen bekend zouden zijn en om die reden niet of nauwelijks zouden discrimineren tussen vaardige en minder vaardige leerlingen.
- Voorafgaand aan de eigenlijke woordselectie, hebben we voor elke groep een range van geometrische gemiddelden vastgesteld gebaseerd op Schrooten en Vermeer. Het geometrisch gemiddelde laat zien of een woord(combinatie) in veel of weinig situaties voorkomt. Zo hebben woorden die in één methode zeer frequent voorkomen, maar niet in het taalaanbod van de leerkracht of in jeugdboeken, een lager geometrisch gemiddelde dan minder frequente woorden die in meerdere methoden en jeugdboeken en in het taalaanbod van de leerkracht voorkomen.
- Uitgaande van deze ranges hebben we voor groep 7 en 8 woorden en vaste combinaties van woorden geselecteerd, waarbij we hebben gelet op een evenwichtige spreiding in geometrische gemiddelden. Voor elke range hebben we gestreefd naar een min of meer gelijke verdeling van woorden met een 'hoog', 'gemiddeld' en 'laag' geometrisch gemiddelde om leerlingen niet onnodig te confronteren met alleen gemakkelijke of moeilijke woorden.
- Om een zo breed mogelijk scala aan woorden te kunnen garanderen, was ook een evenwichtige verdeling in woordsoorten het streven. Hierbij zijn we uitgegaan van een verdeling in woordsoorten zoals die in de Nederlandse taal voorkomt. Zelfstandig naamwoorden zijn in de meerderheid, gevolgd door werkwoorden en bijvoeglijke naamwoorden. Functiewoorden komen daarentegen beduidend minder voor.
- Vervolgens konden er aan de hand van de geselecteerde stimuluswoorden opgaven geconstrueerd worden. Tijdens het proces van opgavenconstructie werden de geometrisch gemiddelde waarden van de woorden in de antwoordalternatieven bekeken. Bleken deze lager dan het geometrisch gemiddelde behorende bij het stimuluswoord dan werd het betreffende antwoordalternatief vervangen door een woord met een hoger geometrisch gemiddelde. Dit om uit te sluiten dat de woorden in de antwoord-alternatieven moeilijker waren dan het bevraagde woord c.q. het stimuluswoord. Over het algemeen kan de hoogte van het geometrisch gemiddelde namelijk opgevat worden als graadmeter voor de moeilijkheid van een woord. Hoe hoger het geometrisch gemiddelde des te aannemelijker dat het een relatief eenvoudig woord betreft.

Beschrijving van de stimuluswoorden

Zoals eerder gezegd bevatten de opgaven in de toetsen Woordenschat voor groep 7 en 8 woorden die representatief zijn voor het taalaanbod in deze groepen.

In de toetsen vanaf groep 5 komen geleidelijk aan steeds meer abstracte woorden voor. Zo neemt vanaf groep 5 het 'vakjargon' van het onderwijs een steeds grotere plaats in. Het gaat daarbij om woorden die vrijwel alleen in schoolse contexten geleerd worden en die als het ware over thema's heen aan bod komen (woorden als *methode, bevestigen, doelgericht*). Een enkele keer komen er woorden in de toetsen voor die aansluiten bij een thema of specifiek vakgebied. Deze woorden vormen echter geenszins het uitgangspunt. De toetsen Woordenschat zijn immers methode-overstijgend en sluiten niet aan bij een bepaald thema of vakgebied. Dat dit type woorden zo nu en dan tóch in de toetsen voorkomt, is echter onvermijdelijk.

Wat betreft de woordsoorten zijn de woorden in de opgaven voor groep 7 en 8 op een enkel woord na te karakteriseren als inhoudswoorden. Inhoudswoorden zijn woorden met een duidelijk omschreven betekenis, zoals zelfstandig naamwoorden, werkwoorden en bijvoeglijk naamwoorden, maar ook woorden die een vaste verbinding vormen. Al deze woorden verwijzen naar voorwerpen, ideeën, activiteiten, plaatsen of eigenschappen. Daarnaast zijn er functiewoorden, de zogenaamde 'kleine' woordjes in de taal, die zonder context weinig betekenis hebben en meestal in combinatie met inhoudswoorden voorkomen. Lidwoorden, telwoorden, voornaamwoorden, voorzetsels, voegwoorden en bijwoorden zijn er voorbeelden van.

Tabel 3.7, 3.8 en 3.9 bieden een samenvatting van de overzichten voor respectievelijk de papieren toets van groep 8 en de digitale toetsen van groep 7 en groep 8. In de tabellen zijn de woordsoorten die in de toetsopgaven voorkomen en hun frequentie schematisch weergegeven. De uiteindelijke verdeling van de verschillende woordsoorten in de toetsen weerspiegelt grotendeels de verdeling in woordsoorten in de Nederlandse taal. In de toetsen komen relatief weinig functiewoorden voor en met betrekking tot de inhoudswoorden kunnen we opmerken dat het aantal zelfstandige naamwoorden relatief groot is en direct gevolgd wordt door de werkwoorden en de bijvoeglijke naamwoorden.

Tabel 3.7 Aantal stimuluswoorden ingedeeld naar woordsoort m.b.t. de papieren toets voor groep 8

	Toets B8/M8
Inhoudswoorden	
Zelfstandige naamwoorden	27
Werkwoorden	13
Bijvoeglijke naamwoorden	6
Woorden die een vaste verbinding vormen	10
Functiewoorden	
Bijwoorden	14

Tabel 3.8 Aantal stimuluswoorden ingedeeld naar woordsoort m.b.t. de digitale toetsen voor groep 7

	Toets M7	**Toets E7**
Inhoudswoorden		
Zelfstandige naamwoorden	51	37
Werkwoorden	6	12
Bijvoeglijke naamwoorden	8	10
Woorden die een vaste verbinding vormen	1	6
Functiewoorden		
Bijwoorden	4	5

Tabel 3.9 Aantal stimuluswoorden ingedeeld naar woordsoort m.b.t. de digitale toets voor groep 8

	Toets B8/M8
Inhoudswoorden	
Zelfstandige naamwoorden	26
Werkwoorden	14
Bijvoeglijke naamwoorden	6
Woorden die een vaste verbinding vormen	10
Functiewoorden	
Bijwoorden	14

3.2.3 Selectie van de opgaven

Alle opgaven die in de toetsen Woordenschat zijn opgenomen zijn speciaal voor deze toetsen geconstrueerd door een constructieteam, voornamelijk bestaande uit (oud-)leerkrachten uit het basisonderwijs.
De geconstrueerde opgaven zijn in een landelijk proefonderzoek voorgelegd aan leerlingen in de jaargroepen waarvoor ze bedoeld waren. Daarbij was het streven dat elke opgave door minimaal 300 leerlingen gemaakt zou worden. Het primaire doel van dergelijke onderzoeken is het verkrijgen van informatie over de kwaliteit en de moeilijkheid van de afzonderlijke opgaven. Ook kunnen opgaven met een laag discriminerend vermogen geïdentificeerd en verwijderd worden. Het gaat dan bijvoorbeeld om opgaven die vaker door vaardige leerlingen dan door minder vaardige leerlingen fout beantwoord worden. Daarnaast bieden deze zogenaamde kalibratieonderzoeken de mogelijkheid om aan de deelnemende leerkrachten te vragen of ze inhoudelijke of andersoortige bezwaren hebben tegen woorden of opgaventypen die in de toetsen zijn opgenomen.
De opgaven die psychometrisch geschikt bleken, zijn vervolgens opgenomen in de toetsen ten behoeve van de normeringsonderzoeken. In principe kwamen alle opgaven met een acceptabele moeilijkheid (p-waarde tussen .40 en .90) en een acceptabel discriminerend vermogen (r_{ir} vanaf .20) hiervoor in aanmerking. Echter, naast psychometrische waren ook inhoudelijke criteria bij de opgavenselectie van belang, zie paragraaf 3.2.2: Het proces van woordselectie. Zo wilden we de opgaven zo evenwichtig mogelijk verdelen over de categorieën ‘Betekenis’ en ‘Betekenisrelaties’ én over de verschillende opgaventypen heen. Ook zochten we naar een zekere balans in woordsoorten. Om de toetsen overzichtelijk te houden, wilden we bovendien opgaven behorend tot eenzelfde opgaventype clusteren en streefden we naar minstens twee opgaven per opgaventype. Verder probeerden we te vermijden dat opgaven die elkaar ‘bijten’ (bijvoorbeeld omdat in de antwoordalternatieven hetzelfde woord voorkomt) in hetzelfde toetsdeel geplaatst zouden worden.
Met het oog op eventuele uitval van opgaven wegens ‘zwak’ functioneren, zijn in de normeringsonderzoeken meer opgaven opgenomen dan het aantal dat was voorzien voor de definitieve toets. De kans op zwak functionerende opgaven in de normeringsonderzoeken was overigens niet heel groot, omdat de ‘zwakke’ opgaven al verwijderd waren op basis van de resultaten van het kalibratieonderzoek.
Van alle opgaven die zijn meegegaan in het normeringsonderzoek zijn de gekalibreerde p-waarde en de r_{ir} bepaald. Bij het samenstellen van de definitieve toetsen zijn vervolgens enkele opgaven verwijderd om te komen tot het gewenste aantal opgaven: 70 opgaven per afnamemoment, zowel voor groep 7 als voor groep 8. Soms vielen er opgaven af die psychometrisch gezien goed functioneerden, maar die op inhoudelijke gronden afgewezen werden (hiervoor werden dezelfde criteria gehanteerd als bij het proefonderzoek). Deze behoorden dan bijvoorbeeld tot een opgaventype dat al voldoende in de toets vertegenwoordigd was. Daarentegen hebben we in een enkel geval ook opgaven gehandhaafd die eigenlijk wat te moeilijk of te gemakkelijk waren of opgaven met een te laag discriminerend vermogen.
De uiteindelijke verdeling van de opgaven was steeds een zo goed mogelijk compromis tussen de eisen op psychometrisch en inhoudelijk gebied en tussen overwegingen van praktische aard om aan het beoogde aantal opgaven te komen.

4 Het normeringsonderzoek

4.1 Opzet en verloop

Met het oog op het ontwikkelen van de toetsen Woordenschat zijn in de periode 2009-2012 opgaven geconstrueerd voor de afnamemomenten medio en eind groep 7 én voor de afnamemomenten begin en medio groep 8. In dezelfde periode zijn deze opgaven in een kalibratieonderzoek voorgelegd aan groepen leerlingen op een groot aantal scholen om gegevens over de kwaliteit en de moeilijkheid van de opgaven te verzamelen. Aansluitend zijn bij een landelijke normgroep referentiegegevens verzameld door de psychometrisch en inhoudelijk meest geschikte opgaven voor te leggen aan leerlingen op de normerings-momenten medio en einde schooljaar groep 7 en begin en medio groep 8.
De normering voor de toetsen M7 en E7 vond in januari en juni 2011 plaats, de normering voor de toets B8/M8 vond respectievelijk plaats in november 2011 (B8) en in januari 2012 (M8). Scholen hebben de mogelijkheid om de toets Woordenschat groep 8 op één van de twee genoemde afnamemomenten aan te bieden.
Op basis van de gegevens van deze onderzoeken konden we vaststellen welke opgaven in de papieren uitgaven Woordenschat voor groep 7 en 8 moesten worden opgenomen. De opgaven uit de (papieren) toetsen voor groep 7 en 8 zijn vervolgens ingezet in een kalibratieonderzoek papier-digitaal, een vergelijkend onderzoek dat halverwege en aan het einde van het schooljaar plaatsvond in groep 7 en aan het begin van het schooljaar in groep 8. Daarna konden we bepalen of deze opgaven geschikt waren om te worden opgenomen in de digitale toetsen Woordenschat.

Kalibratieonderzoek papieren opgaven
Eerder merkten we al op dat in het kalibratieonderzoek dat aan de opgavenbanken ten grondslag ligt, is uitgegaan van een onvolledig design: niet alle leerlingen in de steekproef van het kalibratieonderzoek hebben alle opgaven gemaakt. De opgaven zijn verdeeld over clusters en aan elke leerling zijn een of meer opgavenclusters voorgelegd. Clusters die gezamenlijk aan een groep leerlingen zijn voorgelegd, worden 'boekjes' of ‘booklets’ genoemd. De verschillende boekjes overlappen elkaar. Deze overlap zorgt ervoor dat het design verbonden is. Dit is een noodzakelijke voorwaarde om de conditioneel meest aannemelijke schattingen van de itemparameters (CML-estimates) te kunnen bepalen. Een voorbeeld van zo’n design staat in de verantwoording van de Toetsen Begrijpend Lezen (Staphorsius, Krom, Kleintjes en Verhelst, 2001).
In januari 2010 is een kalibratieonderzoek uitgevoerd voor de afnamemomenten M7 en E7. Informatie over dit onderzoek is te vinden in de Wetenschappelijke verantwoording Woordenschat groep 5 tot en met 7 (Van Berkel e.a., 2012). In januari 2011 heeft een kalibratieonderzoek plaatsgevonden voor de afname-momenten B8 en M8. Dit kalibratieonderzoek is uitgevoerd met een overlap naar eerdere kalibratie- en normeringsonderzoeken. Er zijn 151 opgaven in vier verschillende boekjes geproeftoetst. Het betrof 964 leerlingen met een gemiddeld aantal leerlingantwoorden van 470 per opgave. Ook de woordenschat-opgaven uit de Entreetoets voor groep 7 zijn in dit onderzoek meegenomen en konden op de vaardigheids-schaal LOVS-Woordenschat gekalibreerd worden.
De genoemde kalibratieonderzoeken hadden alle tot doel te onderzoeken of de woordenschatopgaven op de op dat moment beschikbare schaal LOVS-Woordenschat pasten. De inhoudelijk én psychometrisch meest geschikte opgaven zijn vervolgens ingezet in een normeringsonderzoek, waarna de waarden van de itemparameters zijn bepaald. Hoewel de kalibratieonderzoeken uiteindelijk van gering belang zijn, zijn ze hier voor de volledigheid toch besproken.

Normeringsonderzoek papieren opgaven

Het normeringsonderzoek levert aanvullende gegevens op over de kwaliteit en de moeilijkheid van de opgaven en over de landelijke verdeling van de vaardigheid van de leerlingen op de verschillende afnamemomenten. De leerlingen in groep 8 zijn zowel op het begin- als op het mediomoment getoetst om in een landelijke normgroep referentiegegevens voor de verschillende afnamemomenten te kunnen verzamelen om op basis daarvan de ontwikkeling van de vaardigheid in woordenschat in kaart te kunnen brengen. De gegevens uit de normeringsonderzoeken zijn samen met die van de kalibratieonderzoeken ingezet om de vaardigheidsverdelingen op de verschillende normeringsmomenten te kunnen bepalen.
De gegevens uit de kalibratieonderzoeken zijn gebruikt om de verschillende (normerings)tijdstippen te verbinden. De normering zelf wordt volledig bepaald door de normeringsonderzoeken.
De resultaten van de leerlingen die aan de normeringsonderzoeken hebben deelgenomen, vormen de basis voor de normeringsgegevens zoals die zijn opgenomen in de handleiding bij de toetsen Woordenschat.
De representativiteit van de steekproeven wordt in paragraaf 4.2 voor elk van de normeringsonderzoeken weergegeven op basis van schoolkenmerken.

Afnamedesign normeringsonderzoek

De afnamedesigns voor de beide normeringsonderzoeken worden weergegeven in tabel 4.1. Voor alle afnamemomenten is een vergelijkbare onderzoeksopzet gebruikt. Het opnemen van een anker met het vorige normeringsmoment is hierbij wenselijk. Hoewel de opgaven in de kalibratieonderzoeken al op één schaal gebracht zijn, willen we in het normeringsonderzoek toch nog een extra zekerheid inbouwen.
De kalibratieonderzoeken hebben immers op een ander tijdstip plaatsgevonden: enkele maanden voor het normeringsonderzoek. De taak M7_Moeilijk bevat de moeilijke opgaven van M7.
Omdat een deel van de leerlingen meerdere keren aan een normeringsonderzoek deelneemt, zijn aan de afname bepaalde restricties verbonden. Het is immers niet wenselijk dat de leerling een bepaalde opgave twee keer maakt. Aan leerlingen die aan het normeringsonderzoek M7 hebben deelgenomen, worden op afnamemoment B8 alleen de taken B8_1 en B8_2 voorgelegd.
In het design is ook het aantal leerlingen vermeld dat aan het normeringsonderzoek heeft deelgenomen. Voor alle normeringsmomenten is het totaal aantal leerlingen ruim voldoende om een verantwoorde normering op te baseren. De taak met opgaven uit M7 wordt hier alleen genoemd om de verbondenheid van het design te verhelderen. De normering van M7-papier is hier niet aan de orde, die is al eerder verantwoord.

Tabel 4.1 Afnamedesigns normeringsonderzoeken B8 en M8 met aantal deelnemende leerlingen (N)

B8

Taak 8_1	Taak 8_2	Taak M7_Moeilijk	N
*		*	150
*	*		1316
	*	*	85

M8

Taak 8_1	Taak 8_2	N
*		127
*	*	584
	*	45

In tabel 4.2 worden de aantallen leerlingen per afnamemoment gegeven. Leerlingen die meer dan 25% van de opgaven niet gemaakt hebben, zijn niet in de berekening van de normeringsgegevens meegenomen. Het is namelijk onduidelijk of deze leerlingen serieus aan de toets gewerkt hebben. Als deze leerlingen in de berekening betrokken zouden worden dan zouden ze de normeringsresultaten kunnen 'vervuilen'.
De uiteindelijke aantallen leerlingen die in de normeringsanalyses zijn opgenomen, staan in de kolom 'Geanalyseerd'.

Tabel 4.2 Aantal leerlingen per afnamemoment

Afnamemoment	Aantal leerlingen	
	Deelgenomen	Geanalyseerd
B8	1560	1551
M8	756	755

Kalibratieonderzoek papier-digitaal
Door middel van een kalibratieonderzoek papier-digitaal is nagegaan of de gedigitaliseerde opgaven uit de papieren toetsen voor groep 7 en 8 op de schaal Woordenschat passen. Het afnamedesign zoals gehanteerd voor de kalibratie van de digitale opgaven is te vinden in tabel 4.3. Voor de afnamemomenten M7, E7, B8/M8 is steeds eenzelfde onderzoeksopzet gehanteerd.

Tabel 4.3 Afnamedesign kalibratieonderzoek papier-digitaal

Boekje	P1	P2	P3	P4	D1	D2	D3	D4
1	*	*					*	*
2		*	*		*			*
3			*	*	*	*		
4	*			*		*	*	

De delen P1, P2, P3 en P4 bevatten elk (ongeveer) een kwart van de opgaven uit de reeds verschenen papieren toetsen Woordenschat. De delen D1, D2, D3 en D4 zijn de digitale varianten van de papieren delen P1, P2, P3 en P4. Door de papieren en digitale delen te combineren, maakt elke leerling die aan het onderzoek papier-digitaal deelneemt een toets, waarbij de ene helft bestaat uit papieren en de andere helft uit digitale opgaven. De leerlingen krijgen van elke opgave slechts één variant voorgelegd. Dat houdt in dat ze een opgave óf op papier óf digitaal aangeboden krijgen. De digitale delen zijn ten opzichte van de papieren delen met enkele opgaven uitgebreid om bij de samenstelling van de uiteindelijke digitale toetsen enige vrijheid in de opgavenselectie te hebben. De uiteindelijke digitale toetsen bevatten daarmee niet noodzakelijkerwijs exact dezelfde opgaven als de uiteindelijke papieren toetsen.
In tabel 4.4 worden de aantallen leerlingen per afnametijdstip voor het onderzoek papier-digitaal weergegeven. Merk op dat de verzamelde data uit het onderzoek papier-digitaal zijn toegevoegd aan de dataset die dient voor de schaling van de opgaven. Omdat voor deze opgaven zowel in het kalibratie- als in het normeringsonderzoek gegevens verzameld zijn, is het (totale) aantal leerlingantwoorden voor de papieren opgaven veel hoger dan vermeld in tabel 4.4. Elke digitale opgave is op afnamemoment M7 door gemiddeld 284, op afnamemoment E7 door gemiddeld 389 en op afnamemoment B8 door gemiddeld 1461 leerlingen gemaakt. Aangezien het hier een kalibratieonderzoek en geen normeringsonderzoek betreft, is het niet noodzakelijk om de digitale opgaven ook op moment M8 te toetsen. Het doel is immers om de digitale opgaven op de vaardigheidsschaal Woordenschat te passen en daarvoor is een papier-digitaal onderzoek voldoende.

Tabel 4.4 Aantal leerlingen in het kalibratieonderzoek papier-digitaal

Boekje	M7	E7	B8
1	131	193	657
2	132	178	768
3	158	189	710
4	145	180	715

Onderstaande tabellen laten de resultaten van de kalibratie van de onderzoeken papier-digitaal zien.
In tabel 4.5 staan de resultaten van de R1c-toets. Deze toets kan worden opgevat als een overall-toets van de totale kalibratie (zie Verhelst, Glas en Verstralen, 1995).

Tabel 4.5 R1c-toets per onderzoek papier-digitaal

	R1c	df	P
M7	2681	1981	0,000
E7	2745	1982	0,000
B8	4170	2512	0,000

Deze R1c-toets is asymptotisch (dat wil zeggen voor een heel groot aantal leerlingen) chi-kwadraat verdeeld; het aantal vrijheidsgraden is te vinden in de kolom 'df' terwijl de overschrijdingskans in kolom 'p' staat. Voor alle onderzoeken papier-digitaal kunnen we op grond van tabel 4.5 concluderen dat de kalibratie als geslaagd mag worden opgevat. Als vuistregel daarvoor kan worden aangehouden dat de R1c-waarde niet significant is en, in gevallen waarin dat wel zo is, niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden. Voor de afnamemomenten M7 en E7 wordt aan deze voorwaarde ruimschoots voldaan. Voor afnamemoment B8 wordt aan deze voorwaarde net niet voldaan; de reden ligt in het feit dat het aantal deelnemende leerlingen op dit moment veel groter was.
In het COTAN Beoordelingssysteem (COTAN 2010, p. 40) wordt ook nog een andere methode genoemd om de modelpassing te verantwoorden. 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 opgave 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 is aan 0,20. Waarden tussen 0,30 en 0,40 kunnen nog als voldoende worden beschouwd. Voor de onderzoeken papier-digitaal worden de waarden weergegeven in tabel 4.6.

Tabel 4.6 Nauwkeurigheid van de itemparameterschattingen (constante 'c')

Toetsmoment	Range	Gemiddelde
M7	0,095-0,643	0,207
E7	0,034-0,345	0,094
B8	0,051-0,203	0,083

Voor alle onderzoeken papier-digitaal geldt dat de gemiddelde 'c' als goed moet worden beoordeeld. Bovendien geldt dat alle c-waarden, behalve vier items bij het afnamemoment M7, minimaal als voldoende aangemerkt kunnen worden.

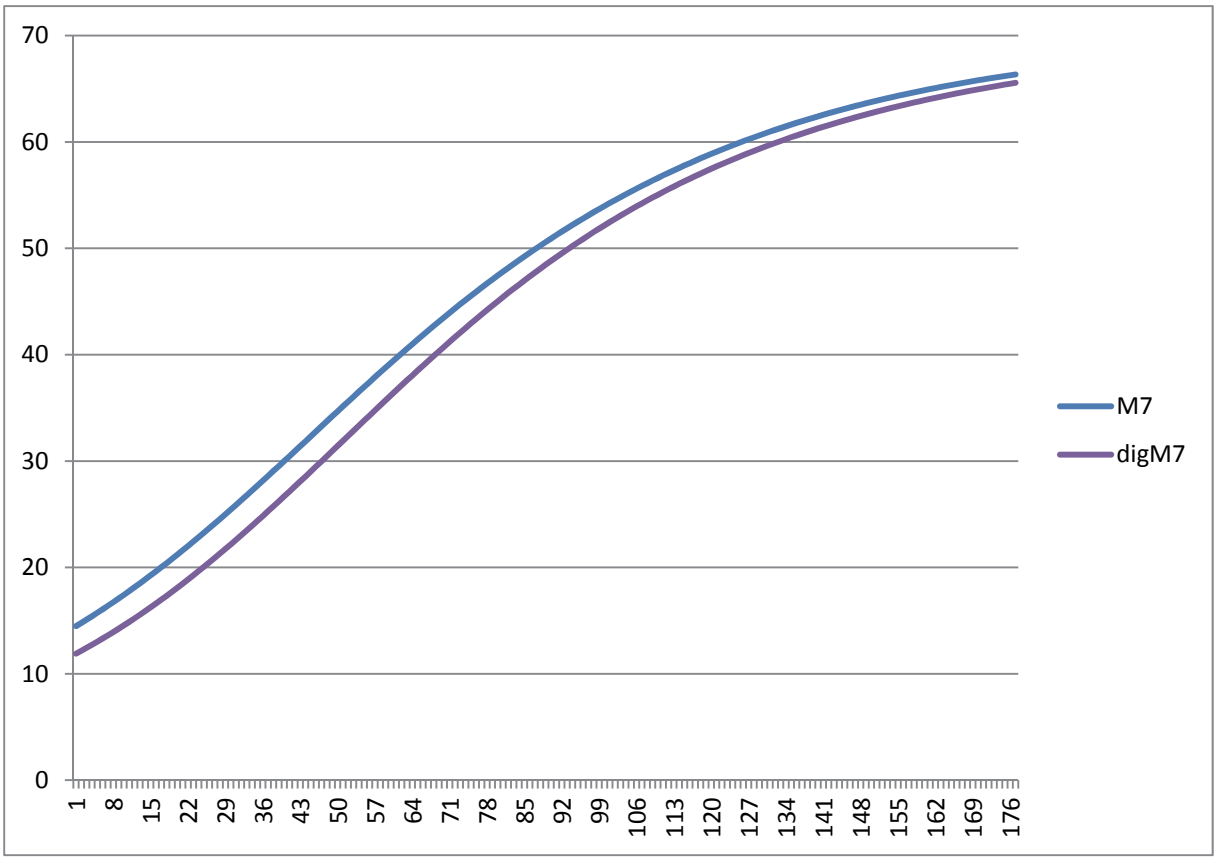
Nadat de kalibratie voor de verschillende onderzoeken papier-digitaal was afgerond, vond het samenstellen van de digitale toetsen plaats. Hierin zijn zowel inhoudelijke als psychometrische aspecten meegenomen. Voor alle afnamemomenten zijn de resultaten m.b.t. de betrouwbaarheid en de meetnauwkeurigheid te vinden in hoofdstuk 5. Op deze plaats willen we alleen opmerken dat alle digitale toetsen wat betreft betrouwbaarheid en meetnauwkeurigheid als goed tot zeer goed aangemerkt kunnen worden.
Voor elk afnamemoment is zowel een papieren als een digitale toets beschikbaar (die beide bestaan uit 70 opgaven). Voor alle afnamemomenten geldt dat ongeveer 95% van de digitale toets gedigitaliseerde papieren opgaven bevat; de overige opgaven zijn alleen digitaal afgenomen. Een terechte vraag die zich hier aandient is of leerlingen niet systematisch bevoor- of benadeeld worden als ze in plaats van een papieren een digitale toets maken. Deze vraag kan worden beantwoord door naar een aantal verschillende aspecten te kijken.
Als eerste vergelijken we voor elk afnamemoment de itemparameters van de papieren en de digitale opgaven. Daartoe vergelijken we eerst de gekalibreerde p-waarden van de papieren en de digitale opgaven. Dit zijn de p-waarden die verkregen zouden worden als de normeringspopulatie de betreffende opgaven zou beantwoorden. Alleen de opgaven die zowel in de papieren als in de uiteindelijke digitale toets zijn opgenomen, worden in dit geval met elkaar vergeleken.

Tabel 4.7 Verschil in gekalibreerde p-waarden tussen papieren en digitale opgaven per afnamemoment

	Aantal opgaven	**Minimum**	**Maximum**	**Gemiddelde**
M7	70	-0,121	0,201	0,016
E7	70	-0,094	0,139	0,019
B8	70	-0,041	0,144	0,027

Uit bovenstaande tabel blijkt dat het verschil in gekalibreerde p-waarden tussen de papieren en digitale opgaven klein is. Gemiddeld genomen vallen de gekalibreerde p-waarden van de digitale opgaven iets lager uit dan die van de papieren opgaven, maar dit verschil is verwaarloosbaar. Gezien het feit dat de papieren en de digitale toetsen voor een groot deel ‘dezelfde’ opgaven bevatten, is het aannemelijk dat de beide toetsen vergelijkbaar zijn. Dit laatste kan worden verduidelijkt als we de toetskarakteristieke curves voor de twee toetsen vergelijken. Ter illustratie zijn voor afnamemoment M7 de beide toetskarakteristieke curves weergegeven in figuur 4.1. Voor de afnamemomenten E7 en B8/M8 is sprake van een vergelijkbare afbeelding.

Figuur 4.1 Toetskarakteristieke curves voor de digitale (digM7) en papieren toets (M7) Woordenschat



Uit figuur 4.1 wordt duidelijk dat de papieren en de digitale toetsen voor alle praktische doeleinden als uitwisselbaar verondersteld mogen worden. Voor de leerlingen maakt het geen verschil of zij de papieren dan wel de digitale versie van een toets maken.
Voor de volledigheid willen we hier nog opmerken dat de gegevens uit de hier besproken kalibratie-onderzoeken alleen voor kalibratiedoeleinden gebruikt worden. De normering vindt steeds plaats op basis van de papieren onderzoeken (zie vorige sectie). Aangezien we voor de kalibratie de CML-methode gebruiken, is het niet noodzakelijk dat de steekproef representatief is. Om plafond- en/of bodemeffecten te vermijden, is het echter wel raadzaam om ervoor te zorgen dat de toetsen redelijk ‘op maat’ zijn voor de kalibratiesteekproef. Verdere eisen hoeven er aan de steekproef niet gesteld te worden.

Ten slotte geven we in tabel 4.8 en 4.9 de resultaten van de kalibraties voor elk van de normerings-momenten. Deze zijn bedoeld als aanvulling en extra controle op de eerder gepresenteerde resultaten in tabel 4.5 en 4.6 met betrekking tot het welslagen van de kalibratie.

Tabel 4.8 R1c-toets per afnamemoment

	R1c	Df	P
M7	1041,019	792	0,0002
E7	831,512	956	0,2140
B8/M8	4216,876	2868	0,0000

Tabel 4.9 Nauwkeurigheid van de itemparameterschattingen (constante 'c')

Afnamemoment	Range	Gemiddelde
M7	0,038 - 0,213	0,090
E7	0,031 - 0,197	0,062
B8/M8	0,047 - 0,291	0,094

Ook nu weer blijkt dat de R1c-waarden ruimschoots aan de eisen voldoen. Hetzelfde geldt voor de itemparameterschattingen. We kunnen dus nogmaals concluderen dat de kalibratie geslaagd is, deze keer uitsluitend op basis van de papieren toets op het tijdstip van normeren.

4.2 Representativiteit

In deze paragraaf bespreken we de representativiteit van de normeringssteekproef voor groep 8.
De normering voor de digitale toetsen voor leerjaar 7 is hier niet aan de orde, want deze is af te leiden van de normeringsonderzoeken van de eerder verantwoorde papieren toetsen voor groep 7.
In tabel 4.2 is per afnamemoment het aantal leerlingen weergegeven dat in de normeringsanalyses is opgenomen. De representativiteit van deze onderzoeksgroep wordt onderzocht op basis van de volgende schoolkenmerken: percentage achterstandsleerlingen, geografische spreiding en mate van verstedelijking. Omdat we niet beschikken over de achtergrondkenmerken van de leerlingen die aan de normerings-onderzoeken hebben deelgenomen, maar wél beschikken over achtergrondkenmerken op het niveau van de scholen, beschrijven we de representativiteit van de scholen waar deze leerlingen onderwijs volgen. Hierbij valt nog aan te tekenen dat het voor Cito van praktisch belang is om in de steekproeftrekking de schoolgrootte als relevante variabele mee te nemen om de vereiste steekproefomvang op verantwoorde wijze te kunnen realiseren. De toegepaste steekproeftrekking is een aselecte trekking van scholen, waarbij per school alle leerlingen van de doelgroep in de steekproef zitten. Daarbij bestaat het risico dat de vereiste steekproefgrootte al snel gerealiseerd wordt door deelname van enkele grote scholen. Kleine scholen zouden daardoor mogelijk in de steekproef ondervertegenwoordigd kunnen zijn. Een steekproeftrekking met een vast aantal leerlingen per school stuit op praktische bezwaren van scholen en van Cito.
De school zou dan aselect leerlingen moeten aanwijzen en voor Cito vallen de kosten voor de steekproef aanzienlijk hoger uit, omdat er meer scholen moeten worden geworven. Om problemen met de representativiteit te voorkomen, is er bij de werving van scholen voor gezorgd dat het aantal grote en kleine scholen in de steekproef een representatieve afspiegeling vormt van de verdeling in de populatie.

Percentage achterstandsleerlingen (stratum)
Een belangrijk schoolkenmerk is het percentage achterstandsleerlingen. Om hier zicht op te krijgen heeft Cito het begrip 'stratum' ingevoerd. De strata zijn gedefinieerd op basis van het percentage achterstands-leerlingen (dat wil zeggen leerlingen met een leerlinggewicht ongelijk aan nul). Het leerlinggewicht wordt bepaald aan de hand van het opleidingsniveau van de ouders (zie tabel 4.10a en 4.10b).

Tabel 4.10a Gewichtenregeling op basis van het opleidingsniveau van de ouders

Gewicht	Omschrijving
0.3	Beide ouders of de ouder die belast is met de dagelijkse verzorging heeft opleiding uit categorie 2 gehad
1.2	Eén van de ouders heeft opleiding gehad uit categorie 1 en de ander een opleiding uit categorie 1 óf 2
0	Eén van de ouders of beide ouders hebben opleiding gehad uit categorie 3

Tabel 4.10b Opleidingsniveau ouder

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

In tabel 4.11 is de verdeling van de scholen uit het normeringsonderzoek naar stratum weergegeven. In stratum 1 zitten scholen met een percentage achterstandsleerlingen kleiner dan 10%; voor stratum 2 geldt dat dit percentage ligt tussen 10 en 25% en in het derde stratum ligt dit percentage tussen 25 en 50%. Stratum 4 bevat scholen met meer dan de helft achterstandsleerlingen.

Hier en ook in het vervolg worden alleen die scholen in de vergelijking meegenomen waarvoor de achtergrondkenmerken (in ons geval percentage achterstandsleerlingen) bekend zijn. Een vergelijking van de steekproefverdeling met de landelijke verdeling laat zien dat voor B8 de steekproef wel en voor M8 de steekproef niet representatief genoemd mag worden. Alle terugweegcoëfficiënten zijn immers kleiner dan 2 uitgezonderd voor P>50 bij het afnamemoment M8.

Tabel 4.11 Scholen uit de steekproef verdeeld naar stratum

	Populatie		Steekproef B8		Steekproef M8	
Stratum	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**
1	3728	55,4	29	52,7	16	44,4
2	1975	29,4	15	27,3	12	33,3
3	719	10,7	7	12,7	3	8,3
4	303	4,5	4	7,3	5	13,9
Totaal	6725	100,0	55	100,0	36	100,0

Door de resultaten van de normeringsonderzoeken terug te wegen naar de populatie is er wel een goede schatting mogelijk van de vaardigheid in de populatie. In tabel 4.12 worden de gemiddelde (geobserveerde) vaardigheidsscores in de verschillende strata per afnamemoment weergegeven; deze gemiddelden noemen we de marginale gemiddelden. Merk op dat de marginale gemiddelden aflopen naarmate het percentage achterstandsleerlingen groter wordt.

Tabel 4.12 Marginale gemiddelden (standaarddeviaties) per afnamemoment naar stratum

	Stratum			
	1	**2**	**3**	**4**
B8	96,1 (13,0)	94,5 (13,1)	87,1 (13,5)	83,1 (11,3)
M8	99,7 (13,5)	97,2 (14,3)	93,5 (11,8)	89,2 (12,3)

Met behulp van deze marginale gemiddelden is er een schatting gemaakt van de populatiegemiddelden, waarbij we terugwegen naar de populatie (zie tabel 4.13).

Tabel 4.13 Gemiddelde vaardigheid per afnamemoment voor de steekproef en de populatie (stratum teruggewogen naar populatie)

Afnamemoment	Steekproef	Populatie
B8	94,3	94,1
M8	97,5	97,8

Uit deze tabel wordt duidelijk dat de steekproef en de populatie met betrekking tot de gemiddelde vaardigheid per afnamemoment na weging niet tot nauwelijks verschillen.

Representativiteit naar geografische verdeling
De verdeling van alle scholen in Nederland en van de scholen in de normeringssteekproef naar regio is te vinden in tabel 4.14. Regio Noord bevat de provincies Groningen, Friesland en Drenthe; regio Oost de provincies Overijssel, Gelderland en Flevoland; regio West de provincies Utrecht, Noord-Holland en Zuid-Holland en regio Zuid bestaat uit Noord-Brabant, Limburg en Zeeland. Ook hier zijn er voor alle afname-momenten kleine verschillen tussen de steekproef- en de populatieverdeling te vinden; geen van de terugweegcoëfficiënten is groter dan twee.

Tabel 4.14 Verdeling scholen naar regio per afnamemoment in aantallen en percentages

	Populatie		Steekproef B8		Steekproef M8	
Regio	**Aantal**	**%**	**Aantal**	**%**	**Aantal**	**%**
Noord	1006	15,0	7	12,7	9	25,0
Oost	1658	24,7	16	29,2	6	16,7
West	2557	38,0	15	27,3	10	27,8
Zuid	1504	22,4	17	30,9	11	30,6
Totaal	6725	100,0	55	100,0	36	100,0

Net zoals in de vorige paragraaf hebben we de resultaten van de normeringsonderzoeken teruggewogen naar de populatie. In tabel 4.15 vinden we de marginale gemiddelden in de verschillende regio's per afnamemoment; in tabel 4.16 de teruggewogen gemiddelden.

Tabel 4.15 Marginale gemiddelden (standaarddeviaties) per afnamemoment naar regio

	Regio			
	Noord	**Oost**	**West**	**Zuid**
B8	94,2 (12,8)	95,0 (13,0)	94,9 (13,9)	92,7 (13,2)
M8	94,3 (13,8)	97,8 (15,9)	99,2 (14,3)	96,2 (12,7)

Tabel 4.16 Gemiddelde vaardigheid per afnamemoment: regio teruggewogen naar populatie

	Steekproef	Populatie
B8	94,3	94,4
M8	97,5	97,7

Vergelijken we de populatie- en de steekproefgegevens in tabel 4.16 dan zien we dat er vrijwel geen verschillen in vaardigheid zijn. Voor zover er in de steekproef dus sprake is van afwijkingen van de populatie in de verdeling naar regio, kan geconcludeerd worden dat deze afwijkingen geen noemenswaardig effect hebben.

Representativiteit naar verstedelijking
De verdeling naar verstedelijking van alle scholen en van de scholen in de normeringssteekproeven wordt weergegeven in tabel 4.17 (aantallen) en tabel 4.17a (percentages). In de steekproef van B8 zijn de niet verstedelijkte gebieden enigszins ondervertegenwoordigd. In de steekproef van M8 zijn de sterk verstedelijkte en de weinig verstedelijkte gebieden enigszins ondervertegenwoordigd, terwijl de matig verstedelijkte gebieden oververtegenwoordigd zijn.

Tabel 4.17 Aantal scholen naar verstedelijking per afnamemoment in aantallen

Mate van	Landelijk		Steekproef	
verstedelijking	Aantal	%	B8	M8
zeer sterk	863	12,8	8	4
sterk	1478	22,0	14	6
matig	1308	19,4	10	12
weinig	1889	28,1	16	6
geen	1187	17,7	7	8
totaal	6725	100,0	55	36

Tabel 4.17a Aantal scholen naar verstedelijking per afnamemoment in percentages

Mate van	Landelijk	Steekproef	
verstedelijking		B8	M8
zeer sterk	12,8	14,5	11,1
sterk	22,0	25,5	16,7
matig	19,4	18,2	33,3
weinig	28,1	29,1	16,7
geen	17,7	12,7	22,2
totaal	100,0	100,0	100,0

De marginale gemiddelden voor de vaardigheidsscore op de afnamemomenten B8 en M8 voor de mate van verstedelijking staan vermeld in tabel 4.18.

Tabel 4.18 Marginale gemiddelden (standaarddeviaties) per afnamemoment naar verstedelijking

	Verstedelijking				
	Zeer sterk	Sterk	Matig	Weinig	Geen
B8	94,8 (14,5)	93,9 (13,7)	94,8 (12,4)	93,6 (13,0)	95,0 (12,2)
M8	94,0 (13,7)	99,9 (12,8)	96,7 (14,1)	96,8 (14,7)	97,8 (14,0)

Met deze marginale gemiddelden is een schatting gemaakt van de populatiegemiddelden, waarbij er is teruggewogen naar de populatie, zoals dat ook voor stratum en regio is gebeurd.

Tabel 4.19 Gemiddelde vaardigheid per afnamemoment: verstedelijking teruggewogen naar populatie

Afnamemoment	Steekproef	Populatie
B8	94,3	94,3
M8	97,5	97,3

Vergelijken we de vaardigheid per afnamemoment voor de verstedelijking teruggewogen naar de populatie met de steekproefgegevens (tabel 4.19) dan zien we dat er vrijwel geen verschillen in vaardigheid zijn. Voor zover er in de steekproef dus sprake is van afwijkingen van de populatie in de verdeling naar mate van verstedelijking, kan geconcludeerd worden dat deze afwijkingen geen noemenswaardig effect hebben.

Representativiteit naar sekse
Voor de normering is het van belang dat alle leerlingen in een jaargroep vertegenwoordigd zijn. Aangezien er in Nederland geen aparte jongens- en meisjesscholen zijn nemen wij aan – gegeven de wijze van steekproeftrekking – dat er sprake is van een representatieve vertegenwoordiging van jongens en meisjes.

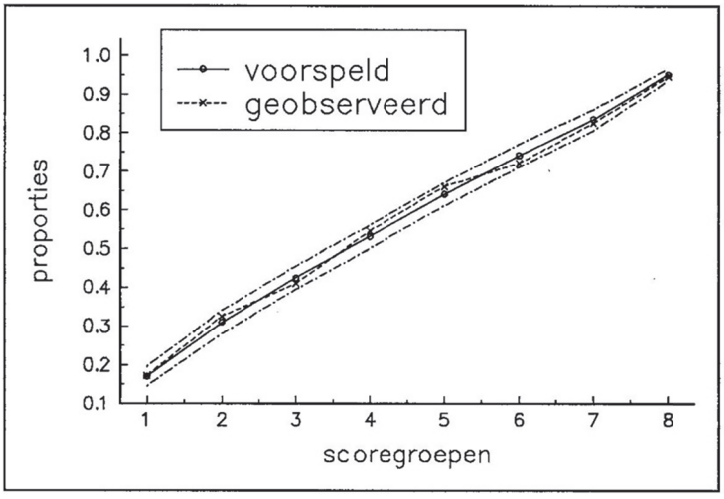
Representativiteit naar leeftijd
Het is voor de normering van belang dat alle leerlingen in een jaargroep vertegenwoordigd zijn. Gegeven de wijze van steekproeftrekking nemen wij aan dat alle leeftijden behorende bij een jaargroep vertegenwoordigd zijn en de verdeling volgen zoals deze in de populatie gebruikelijk is.

4.3 Kalibratie en normering

4.3.1 Toetsing van het IRT-model

In de kalibratieonderzoeken voor de opgavenbanken Woordenschat is steeds getoetst of de opgaven passen bij het model. In deze paragraaf geven we de achtergronden van de toetsing van de opgaven.

Figuur 4.2 Grafische voorstelling van een S_i –toets



De passing van het model illustreren we aan de hand van figuur 4.2 (zie Staphorsius, 1994, blz. 239). Daarin beelden we voor een opgave de gegevens af waarop de zogenaamde S_i -toetsen gebaseerd zijn. Ten behoeve van deze toetsing wordt de totale groep van leerlingen die een verzameling opgaven gemaakt heeft, ingedeeld in een aantal scoregroepen (meestal acht). 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 we op grond van de parameterschattingen kunnen voorspellen. 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_i -toetsings-grootheid (Verhelst, 1994).

Bij de opgaven in onze opgavenbanken hoort een grafische voorstelling van de S_i -toetsing die in grote lijnen met figuur 4.2 overeenkomt. Dit is, zeker gezien de relatief grote aantallen observaties die in het geding zijn, een zeer sterke aanduiding dat het meetinstrument en het meetmodel dat we hebben ontwikkeld en gebruikt, adequaat zijn om het antwoordgedrag van de leerlingen te verklaren. Bovendien blijkt, en dat is vanuit theoretisch oogpunt nog belangrijker, dat de gemeten verschillen in antwoordgedrag tussen de leerlingen te verklaren zijn door één unidimensionaal concept. Tot slot verwijzen we hier nog naar de eerder gepresenteerde analyses, zie tabel 4.8 en 4.9 met betrekking tot de modelpassing in termen van R1c en de nauwkeurigheid van de itemparameterschattingen (‘c’).

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 kalibratieanalyse als puur formeel proces (het analyseren van een grote onvolledige tabel met nullen en enen) 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 opgaven in de opgavenbank meten, dekkend is voor en samenvalt met het construct ‘woordenschat’ zoals in het didactisch en het wetenschappelijk forum wordt bedoeld? De vraag is dan in het geval van woordenschat: kan het unidimensionale concept onder de opgaven in de opgavenbank Woordenschat inderdaad worden opgevat als de vaardigheid in woordenschat? Voor een antwoord op deze vraag verwijzen we naar hoofdstuk 6 Validiteit.

4.3.2 Normering

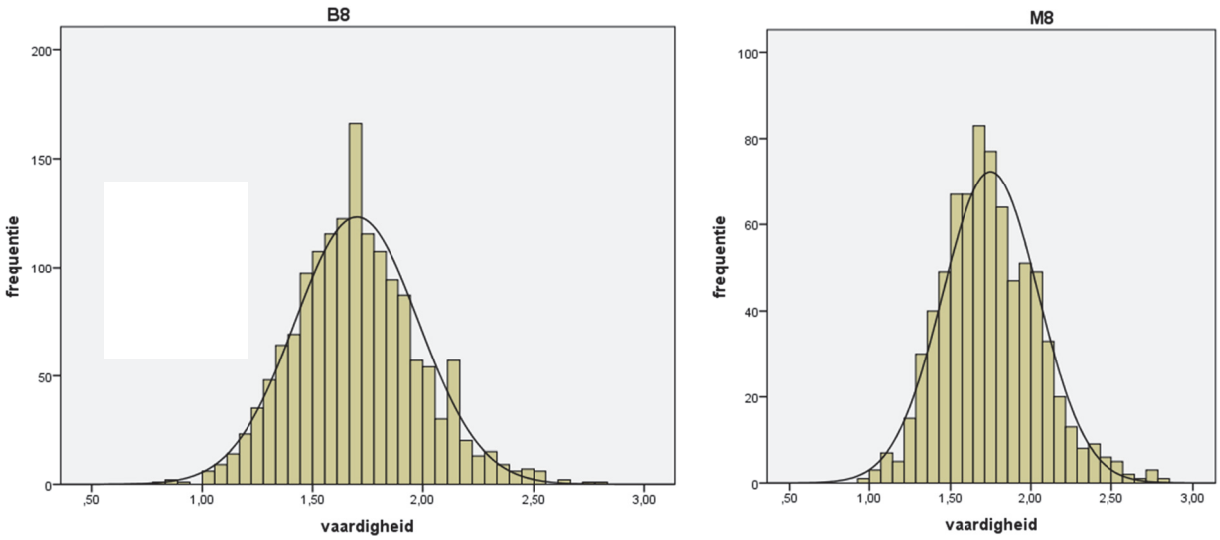
In paragraaf 2.4.2 gaven we belangrijke implicaties voor een gekalibreerde opgavenverzameling.
Het slagen van de kalibratie betekent dat we met een selectie van opgaven uit de opgavenbank de vaardigheid bij een leerling kunnen meten. Hoe nauwkeurig we dat kunnen doen, staat beschreven in paragraaf 5.2.
We kunnen nu een schatting maken van de verdelingen van de vaardigheid in een welomschreven populatie, omdat we selecties van opgaven voorgelegd hebben aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. We schatten het gemiddelde en de standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met deze schattingen kunnen schattingen gemaakt worden van de percentielen in de populatie, die van belang zijn voor de indeling van leerlingen in de niveaugroepen zoals beschreven in paragraaf 2.3.
Een overzicht van de geschatte gemiddelden en de standaardafwijkingen van de vaardigheid op de verschillende normeringsmomenten is te vinden in tabel 4.20. Uit deze tabel blijkt dat de gemiddelde vaardigheid in woordenschat toeneemt, terwijl de spreiding nagenoeg gelijk blijft.

Tabel 4.20 Overzicht van de vaardigheidsverdelingen per afnamemoment

Afnamemoment	Aantal leerlingen	Gemiddelde vaardigheid	Standaardafwijking
B8	1551	94,30	13,55
M8	755	97,45	13,90

Om een indruk te krijgen van de verdeling van de vaardigheid op de verschillende normeringsmomenten is de verdeling van de geobserveerde en de geschatte vaardigheidsverdeling voor elk van de normeringsmomenten grafisch weergegeven in figuur 4.3.

Figuur 4.3 Geobserveerde en geschatte vaardigheidsverdeling voor de normeringsmomenten B8 en M8



Op de horizontale as is de vaardigheid weergegeven; op de verticale as zien we de verdeling van de geobserveerde vaardigheid per afnamemoment. De ingetekende curve is steeds de geschatte vaardigheidsverdeling voor het betreffende afnamemoment. Het zal duidelijk zijn dat de vaardigheid goed beschreven kan worden met een normale verdeling.

We kunnen ook op een andere manier een indruk van de passing krijgen. Daartoe is in tabel 4.21 per afnamemoment de geobserveerde verdeling in de verschillende niveaus weergegeven. De niveaugroepen A, B en C bestrijken elk een kwart van de populatie en het vierde kwartiel is opgesplitst in twee subgroepen: D (15%) en E (10%). De indeling I tot en met V is symmetrisch opgebouwd: vijf niveaugroepen van ieder 20%.

Tabel 4.21 Geobserveerde verdeling van de niveaus per afnamemoment voor de papieren toets B8/M8 (in percentages)

Niveau	B8	M8
E	9,0	10,7
D	13,8	15,6
C	25,5	28,0
B	26,7	19,3
A	24,9	26,3
V	18,3	21,4
IV	18,9	21,8
III	21,9	20,0
II	20,2	15,6
I	20,2	21,2

Voor alle afnamemomenten en voor alle niveaus, met uitzondering van niveau B en niveau II op afnamemoment M8, is te zien dat de afwijking tussen de geobserveerde en de theoretische percentages klein is. Uit bovenstaande blijkt dat de aanname van een normaal verdeelde vaardigheidsverdeling door de data ondersteund wordt. Bovendien is hiervoor ook een toets ontwikkeld, de zogenaamde R0 toets (Verhelst, Glas & Verstralen, 1995). In het onderhavige geval is de p-waarde van deze toets 0,97 (R0=2327,34 met df=2457).

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Betrouwbaarheid

Door de gekozen aanpak in het normeringsonderzoek (i.e. een onvolledig design) is de betrouwbaarheid van de toetsen in klassieke zin niet rechtstreeks te bepalen. Het is echter wel mogelijk om de betrouwbaarheid van iedere toets te schatten door gebruik te maken van het feit dat alle opgaven die zijn opgenomen in de toetsen OPLM-geschaald zijn. Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het gegeven dat de toetsen volledig bestaan uit OPLM-gekalibreerde opgaven. Om relevante beschrijvende gegevens bij de verschillende toetsen te genereren, is gebruikgemaakt 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 duiden we aan met τ(θ). Als we bovendien weten hoe θ in de populatie verdeeld is, kunnen we ook het gemiddelde en de variantie van de ware scores in de populatie bepalen. De variantie van de ware scores in de populatie duiden we aan met het symbool Var(τ). Tussen θ en τ(θ) 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 τ(θ) 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 we met een eenmalige toetsafname niet meer zonder fout de waarde van θ kunnen bepalen. De variantie van de geobserveerde toetsscore duiden we aan met Var(t|τ(θ)), en door weer gebruik te maken van de distributie van θ in de populatie kunnen we ook de gemiddelde variantie van de geobserveerde toetsscores gaan berekenen.

$$Var(t) = E[Var(t \mid \tau(\theta))] \qquad (5.1)$$

Deze variantie kunnen we opvatten als de (gemiddelde) meetfoutvariantie in de metriek van de geobserveerde scores t. In analogie met de theorie over de betrouwbaarheid definiëren we dan

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

waarin MAcc staat voor 'Accuracy of Measurement'.

Tabel 5.1 bevat informatie over de meeteigenschappen van de vaardigheidsschaal Woordenschat.
In de eerste kolom staat het afnamemoment. De maximumscore voor iedere toets is gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. De derde kolom geeft de geschatte gemiddelde scores van de leerlingen op de verschillende toetsen. De vierde kolom bevat informatie over de geschatte standaard-meetfout van iedere toets. De laatste kolom laat zien wat de geschatte betrouwbaarheids- coëfficiënt (MAcc) van de verschillende toetsen is.

De betrouwbaarheidscoëfficiënten zijn zonder uitzondering hoog. Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen Woordenschat) geeft de COTAN aan dat een betrouwbaarheidscoëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 voldoende en een betrouwbaarheidscoëfficiënt hoger dan 0,80 goed (COTAN, 2009, p. 33). Op grond van dit criterium is de meetnauwkeurigheid van alle toetsen goed te noemen.

Tabel 5.1 Beschrijvende gegevens bij de papieren toets Woordenschat B8/M8

Toets	Aantal opgaven	Gemiddelde	Standaard-deviatie	Standaard-meetfout	Betrouwbaarheid
B8	70	44,5	11,4	3,6	0,90
M8	70	46,9	11,2	3,5	0,90

Anders dan bij de papieren toetsen wordt bij de digitale toetsen gewerkt met gewogen scores. Bij de digitale toetsen zijn immers alle antwoorden op de opgaven van een leerling direct beschikbaar. De gewogen score van een leerling verkrijgen we door elke goed beantwoorde opgave te wegen met de discriminatie-index van die opgave en deze gewogen score voor alle opgaven in een toets te sommeren. In tabel 5.2 vindt u de gegevens bij de digitale toetsen Woordenschat. Voor de digitale toets B8/M8 betreft het één toets, die genormeerd is, en dus afgenomen kan worden op twee verschillende momenten, op B8 en M8. Hieruit blijkt dat ook voor de digitale toetsen de betrouwbaarheid hoog is.

Tabel 5.2 Beschrijvende gegevens bij de digitale toetsen Woordenschat M7, E7 en B8/M8

Toets	Afname-moment	Aantal opgaven	Max score	Gemiddelde	Standaard-deviatie	Standaard-meetfout	Betrouw-baarheid
M7	M7	70	228	161,8	33,6	10,9	0,90
E7	E7	70	272	184,6	52,6	13,0	0,94
B8/M8	B8	70	227	144,3	38,8	1168	0,91
B8/M8	M8	70	227	152,3	37,9	11,3	0,91

Gezien het feit alle opgaven OPLM-gekalibreerd zijn, kunnen we door middel van een simulatie de test-hertestbetrouwbaarheid schatten. Daarvoor is een dubbele afname gesimuleerd voor een groep van 100.000 leerlingen. Daarbij hebben we de vaardigheidsverdeling van alle leerlingen op elk afnamemoment en de bijbehorende itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselect uit de verdeling genomen en zijn twee onafhankelijke bij deze vaardigheid horende toetsafnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 100.000 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 zouden plaatsvinden. De zo berekende test-hertestbetrouwbaarheden voor de papieren toetsen zijn identiek aan die van tabel 5.1. Voor de digitale toetsen van groep 7 komen de uitkomsten vrijwel exact overeen met de eerder berekende coëfficiënten in tabel 5.2: voor M7 en E7 (0,93 en 0,97). Voor de toets B8/M8 zijn de testhertest betrouwbaarheden voor beide afnamemomenten (B8 en M8) identiek aan die in tabel 5.2.
Deze resultaten leiden dan ook tot dezelfde conclusies wat betreft de betrouwbaarheid van de toetsen Woordenschat.

5.2 Meetnauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen Woordenschat en geven geen beeld van de lokale meet-nauwkeurigheid. De betrouwbaarheidstabellen 5.3 en 5.4 doen dat wel.
Zo laat tabel 5.4 bijvoorbeeld zien dat 76% van de leerlingen die bij de M7-toets in scoregroep E vallen met

zowel hun geschatte vaardigheidsscore als met hun werkelijke vaardigheidsscore. Anders gezegd: de kans dat een E-leerling terecht als een E-leerling wordt bestempeld, is 76%. Verder laat de tabel zien dat 23% van de leerlingen in niveaugroep E een vaardigheidsscore heeft die in werkelijkheid in (de aangrenzende) scoregroep D valt.

Verdere gedetailleerde informatie over de meetnauwkeurigheid van de toetsen is te vinden in de handleidingen van de toetspakketten Woordenschat (Cito, 2011, 2012). Zie hiervoor de schaalscoretabellen in bijlage 2 van de handleidingen waarbij in de laatste kolom het score-interval vermeld is. In deze kolom staat voor iedere ruwe score op elke toets het 67-procents-betrouwbaarheidsinterval voor de bijbehorende vaardigheidsschatting.

Tabel 5.3 Betrouwbaarheidstabellen bij de papieren toets Woordenschat per afnamemoment

Toets B8/M8 Afnamemoment B8						**Toets B8/M8 Afnamemoment B8**					
	Scoregroepen E tot en met A						Scoregroepen V tot en met I				
Scoregroep waarin ware score valt	E	D	C	B	A	Scoregroep waarin ware score valt	V	IV	III	II	I
E	75,1	9,9	0,1	0,0	0,0		80,7	13,2	0,5	0,0	0,0
D	24,2	61,6	11,7	0,1	0,0		18,7	62,9	22,4	1,6	0,0
C	0,7	28,1	64,6	16,0	0,3		0,6	22,6	55,0	24,4	1,3
B	0,0	0,4	23,1	64,0	16,3		0,0	1,2	21,4	56,7	20,7
A	0,0	0,0	0,5	19,8	83,4		0,0	0,0	0,8	17,3	78,0

Toets B8/M8 Afnamemoment M8						**Toets B8/M8 Afnamemoment M8**					
	Scoregroepen E tot en met A						Scoregroepen V tot en met I				
Scoregroep waarin ware score valt	E	D	C	B	A	Scoregroep waarin ware score valt	V	IV	III	II	I
E	79,2	10,9	0,1	0,0	0,0		80,0	12,1	0,4	0,0	0,0
D	20,4	61,3	12,1	0,2	0,0		19,3	62,1	19,9	1,4	0,0
C	0,4	27,5	66,2	18,9	0,5		0,7	24,2	54,2	22,9	1,5
B	0,0	0,4	21,2	62,5	17,5		0,0	1,6	24,4	56,8	21,5
A	0,0	0,0	0,4	18,5	82,0		0,0	0,0	1,2	19,0	77,0

Tabel 5.4 Betrouwbaarheidstabellen bij de digitale toetsen Woordenschat

Toets M7						**Toets M7**					
	Scoregroepen E tot en met A						Scoregroepen V tot en met I				
Scoregroep waarin ware score valt	E	D	C	B	A	Scoregroep waarin ware score valt	V	IV	III	II	I
E	75,9	9,9	0,1	0,0	0,0		81,5	13,0	0,4	0,0	0,0
D	23,4	61,0	12,8	0,3	0,0		17,9	59,1	19,2	1,8	0,0
C	0,7	28,6	63,8	20,0	0,8		0,6	25,5	51,8	22,9	2,0
B	0,0	0,5	22,6	61,2	20,2		0,0	2,4	26,8	54,6	21,9
A	0,0	0,0	0,6	18,6	79,0		0,0	0,0	1,8	20,6	76,1

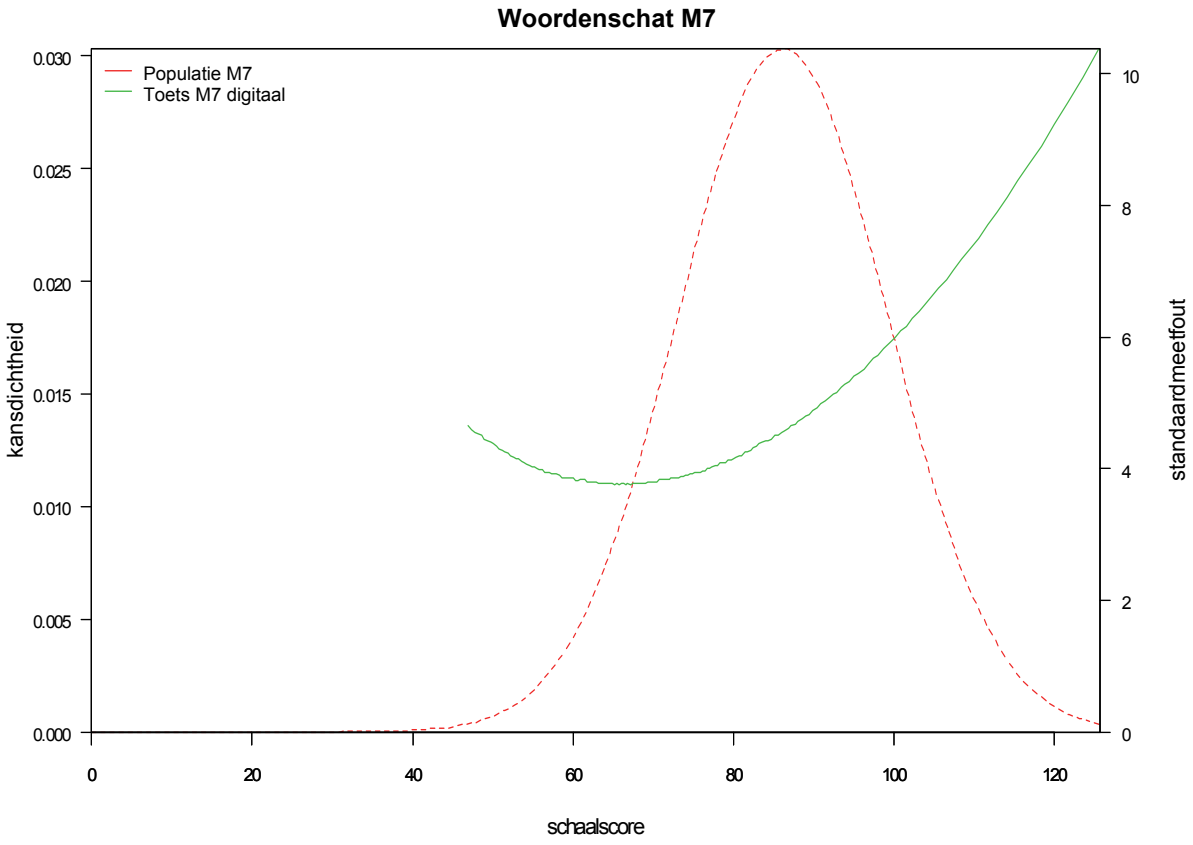
Toets E7						**Toets E7**					
	Scoregroepen E tot en met A						Scoregroepen V tot en met I				
Scoregroep waarin ware score valt	E	D	C	B	A	Scoregroep waarin ware score valt	V	IV	III	II	I
E	83,2	8,8	0,0	0,0	0,0		86,7	10,6	0,0	0,0	0,0
D	16,8	70,4	10,0	0,0	0,0		13,3	69,3	16,3	0,4	0,0
C	0,0	20,8	73,5	14,8	0,1		0,1	19,6	62,8	18,9	0,3
B	0,0	0,0	16,4	69,9	14,2		0,0	0,4	20,6	63,7	16,5
A	0,0	0,0	0,1	15,3	85,7		0,0	0,0	0,3	17,0	83,1

Toets B8						**Toets B8**					
	Scoregroepen E tot en met A						Scoregroepen V tot en met I				
Scoregroep waarin ware score valt	E	D	C	B	A	Scoregroep waarin ware score valt	V	IV	III	II	I
E	77,9	11,1	0,1	0,0	0,0		82,1	12,9	0,3	0,0	0,0
D	21,6	62,9	12,4	0,1	0,0		17,5	62,8	19,1	1,0	0,0
C	0,5	25,8	67,1	17,8	0,3		0,4	23,2	56,6	21,9	1,0
B	0,0	0,2	20,2	65,9	17,2		0,0	1,1	23,2	59,5	19,5
A	0,0	0,0	0,3	16,2	82,5		0,0	0,0	0,8	17,6	79,5

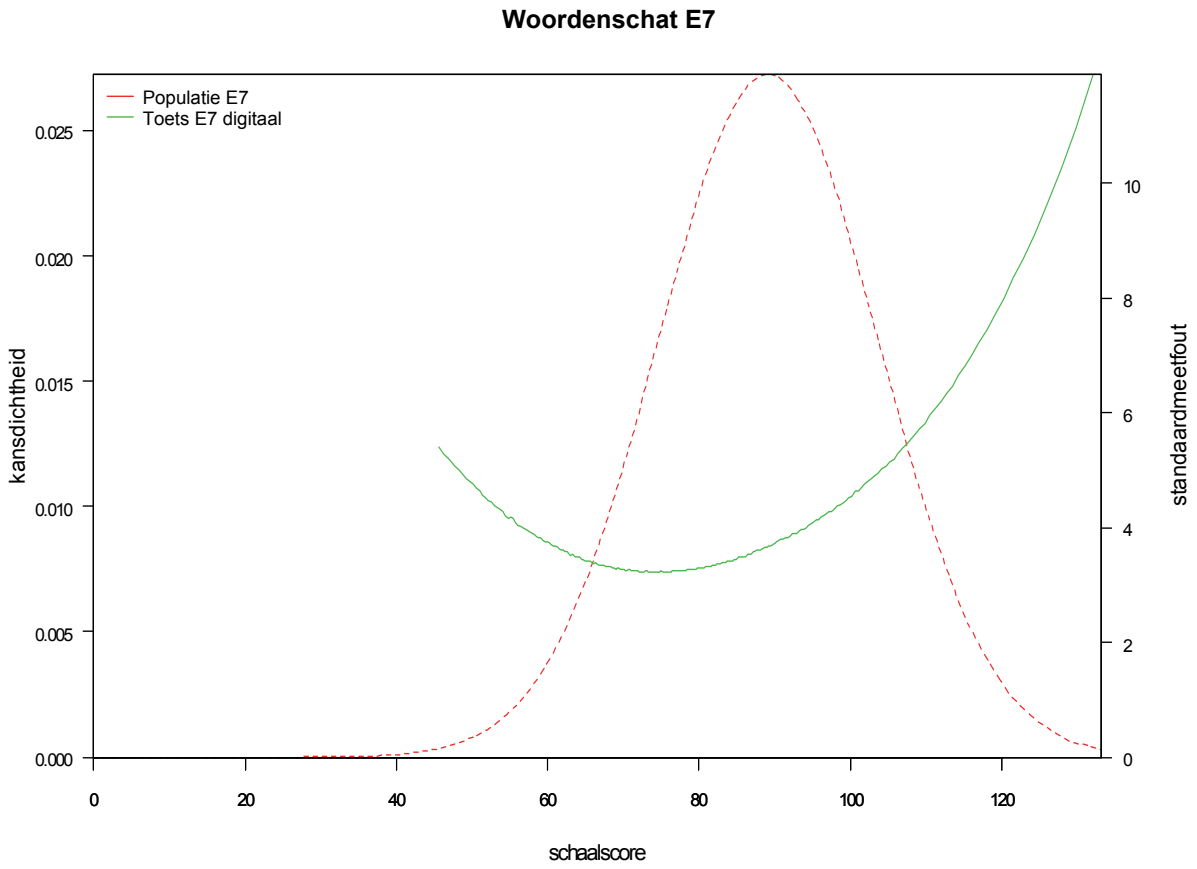
Toets M8						**Toets M8**					
	Scoregroepen E tot en met A						Scoregroepen V tot en met I				
Scoregroep waarin ware score valt	E	D	C	B	A	Scoregroep waarin ware score valt	V	IV	III	II	I
E	77,2	9,8	0,1	0,0	0,0		83,1	12,9	0,2	0,0	0,0
D	22,6	63,3	12,0	0,1	0,0		16,5	62,4	19,1	1,2	0,0
C	0,5	26,7	67,4	17,0	0,3		0,4	23,4	56,2	22,3	1,1
B	0,0	0,3	20,3	65,2	17,0		0,0	1,3	23,7	58,8	20,2
A	0,0	0,0	0,3	17,1	82,7		0,0	0,0	0,8	17,6	78,8

De figuren 5.1 geven nog eens grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de verschillende toetsen. In deze figuren is voor iedere toets de grootte van de meetfout afgebeeld. Ook zijn de kansdichtheidsfuncties voor de normgroepen op de verschillende afnamemomenten opgenomen. Deze laten zien hoe de vaardigheid van de leerlingen verdeeld is over de vaardigheidsschaal in de populatie die de toets gemaakt heeft. De figuren maken duidelijk dat de meetfout kleiner is in de lagere en de gemiddelde vaardigheidsniveaus dan in de hogere vaardigheidsniveaus. Het was de bedoeling dat het discriminerend vermogen van de toets vooral bij de zwakke leerlingen optimaal zou zijn, omdat we met name de vaardigheid van deze leerlingen goed in kaart willen brengen.

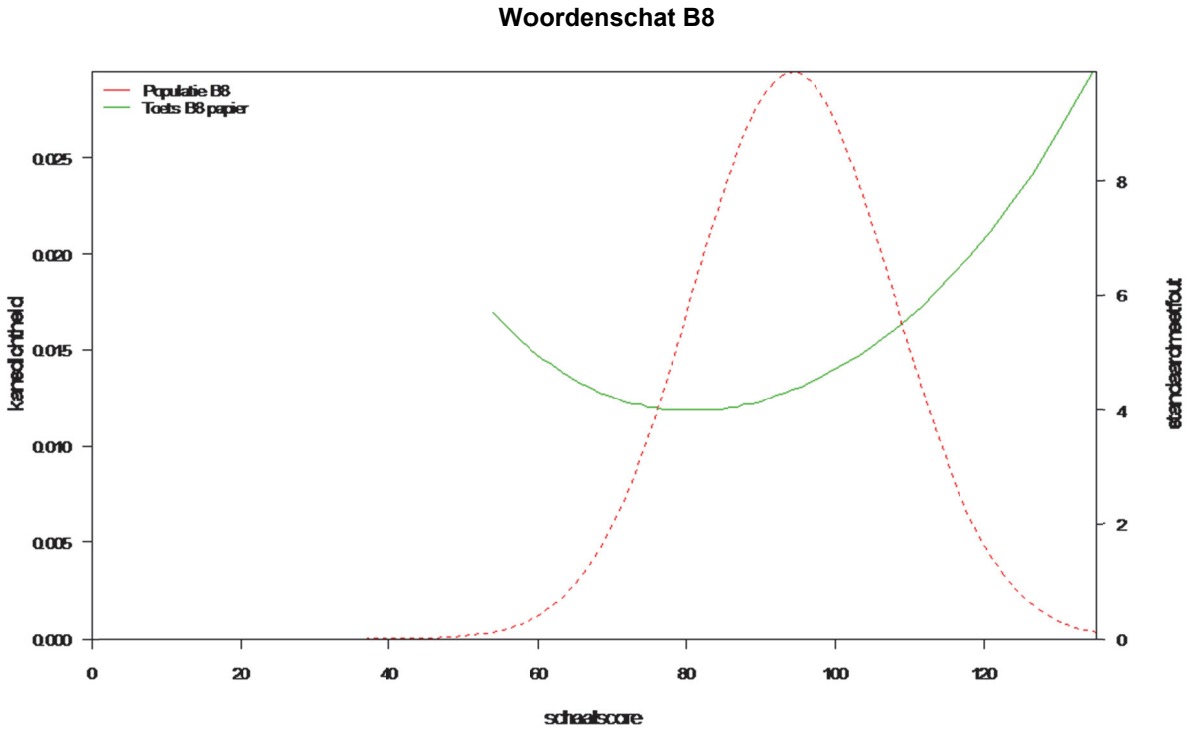
Figuur 5.1a Grootte van de meetfouten voor de digitale toets M7 en de kansdichtheidsfuncties voor de M7-populatie



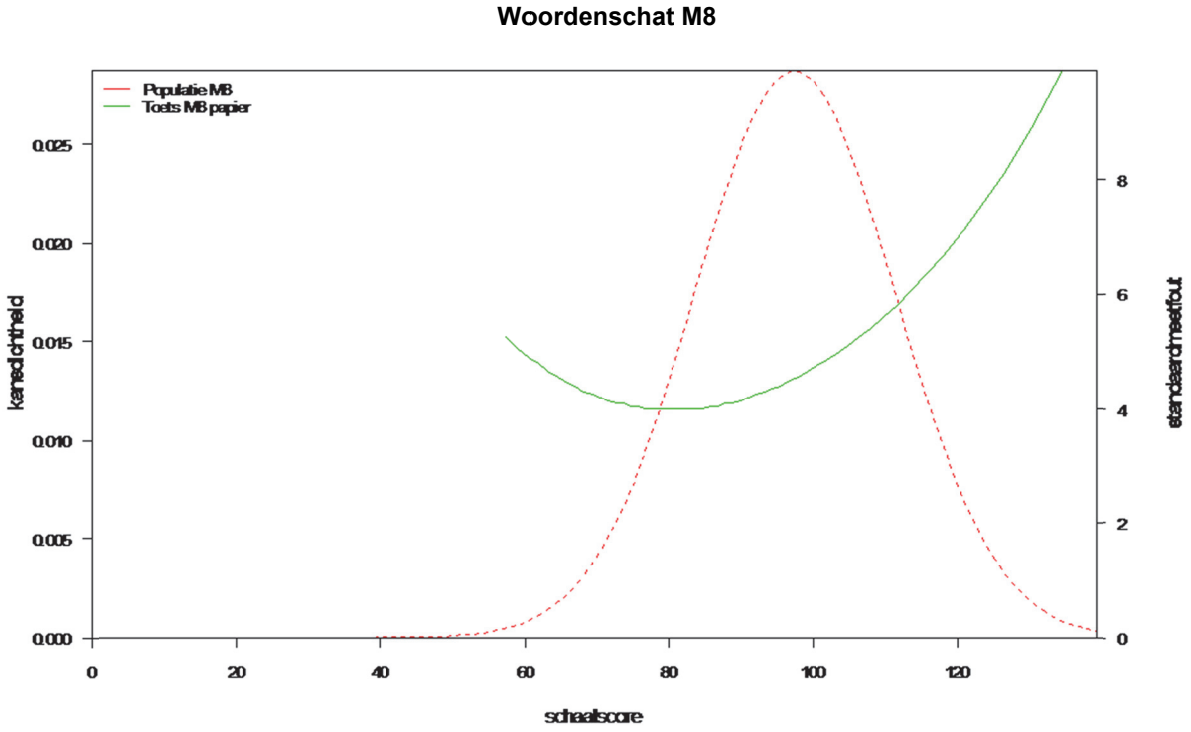
Figuur 5.1b Grootte van de meetfouten voor de digitale toets E7 en de kansdichtheidsfuncties voor de E7-populatie



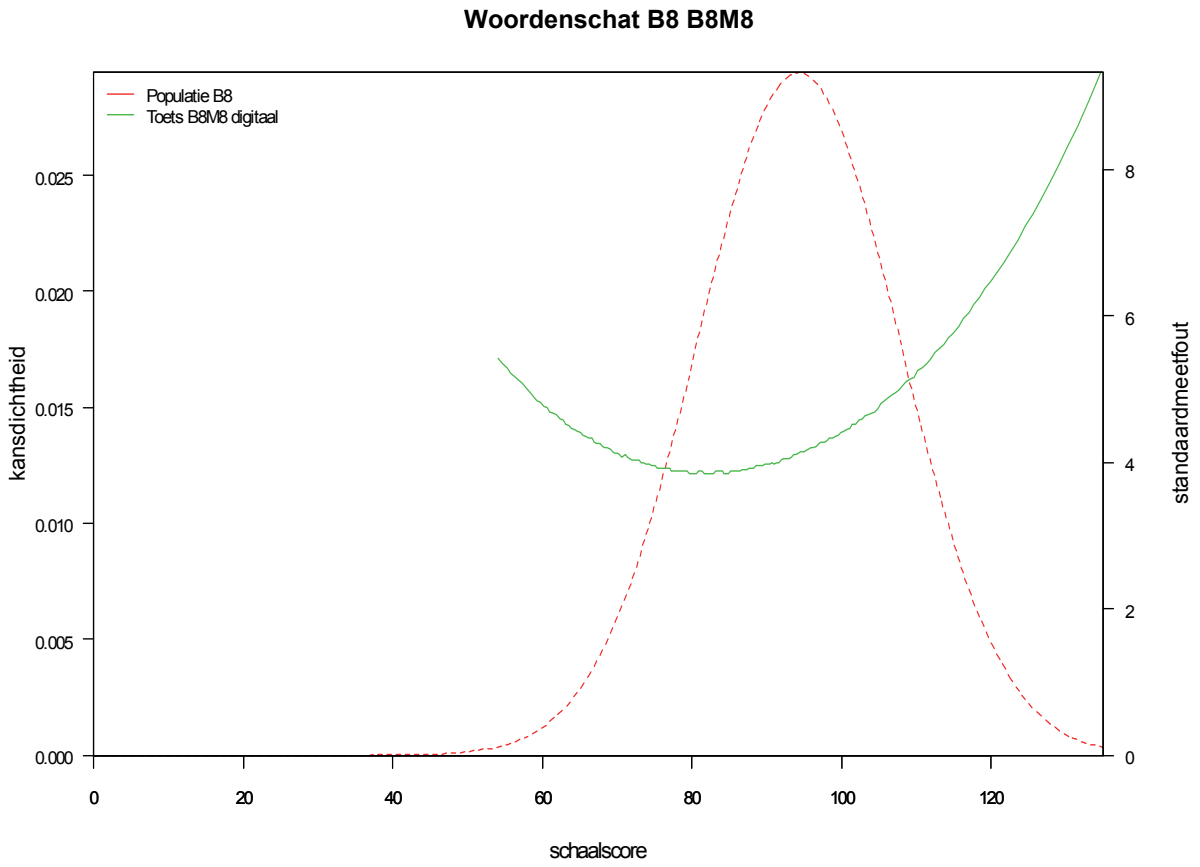
Figuur 5.1c Grootte van de meetfouten voor de papieren toets B8/M8 en de kansdichtheidsfuncties voor de B8-populatie



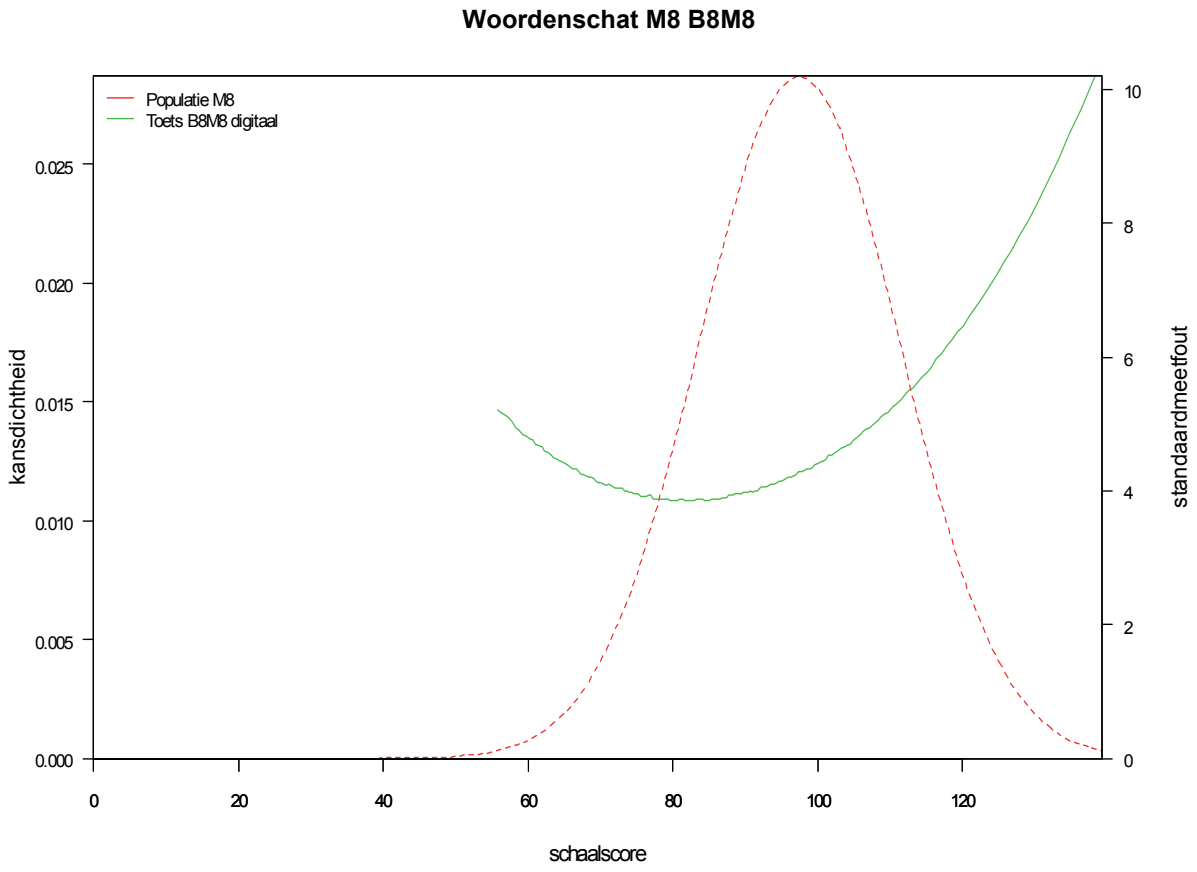
Figuur 5.1d Grootte van de meetfouten voor de papieren toets B8/M8 en de kansdichtheidsfuncties voor de M8-populatie



Figuur 5.1e Grootte van de meetfouten voor de digitale toets B8/M8 en de kansdichtheidsfuncties voor de B8-populatie



Figuur 5.1f Grootte van de meetfouten voor de digitale toets B8/M8 en de kansdichtheidsfuncties voor de M8-populatie



6 Validiteit

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 inhoudsvaliditeit van de toetsen Woordenschat wordt onder meer gegarandeerd door de wijze waarop de opgaven ontwikkeld zijn. In de inhoudsverantwoording die deel uitmaakt van het toetspakket Woordenschat (zie paragraaf 3.2) is al aangegeven dat aan de ontwikkeling van de opgaven een woordfrequentielijst ten grondslag ligt (Schrooten en Vermeer, 1994). Deze lijst maakt duidelijk welke woorden aangeboden worden in de verschillende groepen van het basisonderwijs en welke woorden tijdens het toetsconstructieproces in aanmerking kwamen om in de toetsen opgenomen te worden. Een andere aanwijzing voor de inhoudsvaliditeit is dat ongeveer de helft van de opgaven een beroep doet op aspecten met betrekking tot de breedte en de andere helft van de opgaven op aspecten met betrekking tot de diepte van de woordenschat. Dit sluit nauw aan bij de geraadpleegde literatuur, waaruit naar voren komt dat zowel een brede als een diepe woordenschat een even belangrijke rol vervullen bij het opbouwen van de woordenschat. Hoewel het psychometrisch gezien niet strikt noodzakelijk is om verschillende opgaventypen in de toetsen op te nemen, vonden we dat op basis van de inhoud wél van belang. De verschillende opgaventypen representeren immers de verschillende inhouden, zoals we die in de literatuur en in taalmethoden hebben aangetroffen. Tot slot bevatten de toetsen Woordenschat een representatieve verdeling van woordsoorten uitgaande van de verdeling in woordsoorten zoals die in de Nederlandse taal voorkomt. Hierbij zijn zelfstandig naamwoorden in de meerderheid, gevolgd door werkwoorden en bijvoeglijke naamwoorden. Functiewoorden komen daarentegen beduidend minder voor.

6.2 Begripsvaliditeit

De begripsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de toetsscores toe te schrijven zijn aan verklarende concepten en constructen die deel uitmaken van het theoretische kader dat aan de ontwikkeling van de toets ten grondslag ligt. Hieronder worden vijf aanwijzingen voor de begripsvaliditeit van de toetsen Woordenschat beschreven. Deze hebben betrekking op de passing van het meetmodel (6.2.1), de convergente versus divergente validiteit (6.2.2), de samenhang met leerjaar (6.2.3), de responsiviteit en stabiliteit (6.2.4) en de itemkarakteristieken (6.2.5).

6.2.1 Passing van het meetmodel

De opgaven vormen na de kalibratie een gekalibreerde opgavenbank. Bij de analyse van de leerling-antwoorden is nagegaan of de verschillende opgaven een beroep doen op hetzelfde complex aan vaardigheden. Opgaven die niet voldeden aan de passingscriteria zoals beschreven in paragraaf 4.3.2, zijn uit de opgavenverzameling verwijderd. Het betreft waarschijnlijk opgaven waarop werd gegokt, opgaven die onjuist geformuleerd zijn, opgaven die een slecht onderscheidend vermogen hebben of opgaven die bij nader inzien toch niet alleen woordenschat blijken te meten.
De vraag of het unidimensionale concept onder de opgaven in de opgavenbank Woordenschat kan worden opgevat als de vaardigheid ‘woordenschat’, kan met behulp van de gegevens in hoofdstuk 4 met ‘ja’ beantwoord worden. We hebben verschillende analyses gerapporteerd met betrekking tot de passing van het onderliggende meetmodel van de toetsen, waaruit blijkt dat die passing bevredigend is. De geslaagde kalibratie maakt duidelijk dat het aannemelijk is dat er sprake is van unidimensionaliteit én dat de gekalibreerde opgavenbank de latente trek meet die we de vaardigheid woordenschat noemen.

6.2.2 Convergente/discriminerende validiteit

In het kader van een kwaliteitscontrole van de Eindtoets Basisonderwijs wordt hier een analyse gegeven van de onderzoeksgegevens die in het jaar 2010 verzameld zijn.

Tabel 6.1 Correlaties tussen Woordenschat en de onderdelen Taal (minus Woordenschat), Rekenen-Wiskunde en Studievaardigheden gemeten met de Eindtoets Basisonderwijs (2010)

	Eindtoets Basisonderwijs
Taal (minus Woordenschat)	0,816
Rekenen-Wiskunde	0,591
Studievaardigheden	0,792

De Eindtoets Basisonderwijs (zie de bijbehorende handleiding) bestaat uit drie verschillende onderdelen: Taal, Rekenen-Wiskunde en Studievaardigheden. Elk van deze onderdelen is onderverdeeld in rubrieken. Binnen het onderdeel Taal is Woordenschat een van de rubrieken. De woordenschatopgaven van de Eindtoets Basisonderwijs liggen op dezelfde schaal als de opgaven in de toetsen Woordenschat (zie p. 11). In tabel 6.1 worden de correlatiecoëfficiënten tussen de hierboven genoemde onderdelen uit de Eindtoets Basisonderwijs met de rubriek Woordenschat gerapporteerd. Voor een eerlijke vergelijking gebruiken we bij de berekening van de correlatie tussen Taal en Woordenschat een somscore voor Taal waaruit de score op Woordenschat is verwijderd. Duidelijk is te zien dat de correlatie tussen Woordenschat en Taal groter is dan die tussen Woordenschat en Rekenen-Wiskunde. De correlatie tussen Woordenschat en Studie-vaardigheden is (ook) redelijk groot, maar wel geringer dan die tussen Woordenschat en Taal.

Tabel 6.2 Correlaties tussen Woordenschat en de andere taalrubrieken gemeten met de Eindtoets Basisonderwijs (2010)

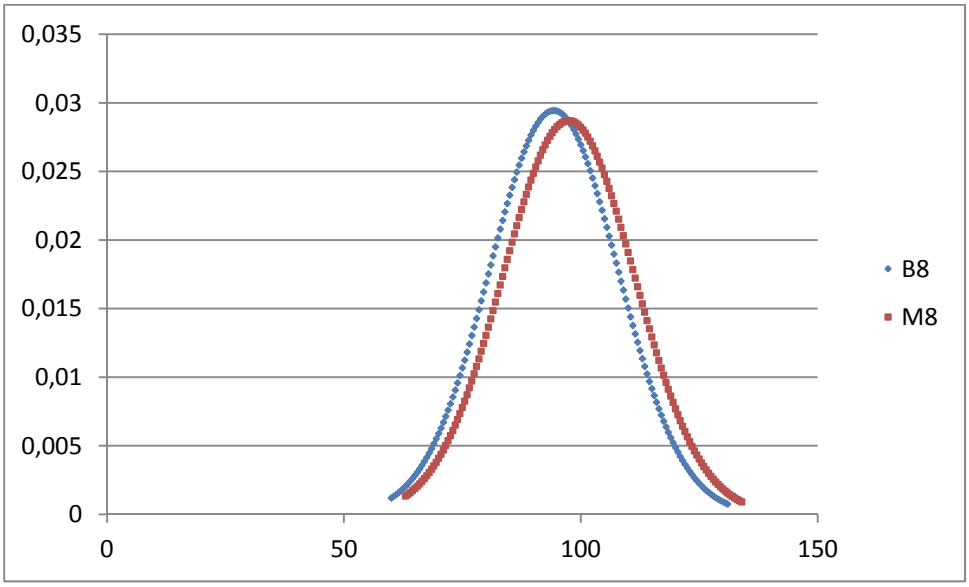
	Eindtoets Basisonderwijs
Schrijfvaardigheid	0,714
Begrijpend lezen	0,849
Spelling	0,520

In tabel 6.2 geven we de correlaties tussen de rubrieken van Taal met Woordenschat weer.
Hier blijkt dat de correlaties tussen Woordenschat en de semantische onderdelen Schrijfvaardigheid en Begrijpend lezen aanzienlijk groter zijn dan tussen Woordenschat en een meer niet-semantische taalaspecten als Spelling. Een aanwijzing voor de validiteit van de woordenschatopgaven zijn de relatief hoge correlaties tussen Begrijpend lezen en Woordenschat en tussen Schrijfvaardigheid en Woordenschat. Woordenschat is dan ook een belangrijke ondersteunende vaardigheid bij deze taalonderdelen.
Wellicht ten overvloede willen we nog vermelden dat alle intercorrelaties binnen leervorderingen in de regel hoog zijn en dat het – qua interpretatie in termen van convergente en divergente validiteit – dus relatieve verschillen in de hoogte van de correlaties betreft.

6.2.3 Samenhang met de variabele leerjaar

In figuur 6.1 staan de vaardigheidsverdelingen voor de afnamemomenten B8 en M8, waarvan de gegevens (gemiddelden en varianties) te vinden zijn in tabel 4.12.

Figuur 6.1 Vaardigheidsverdeling groep 8 per afnamemoment



Uit figuur 6.1 komt naar voren dat de vaardigheid op het gebied van woordenschat licht groeit in de tijd. Deze groei is niet groot omdat er slechts een paar maanden verschil zit tussen de afnamemomenten. De twee afnamemomenten van de toets B8/M8 zijn gesitueerd op dezelfde vaardigheidsschaal Woordenschat en laten een lichte toename van de gemiddelde vaardigheid zien. Deze toename mag verwacht worden op basis van leeftijd, ontwikkeling en de hoeveelheid genoten onderwijs.

6.2.4 Responsiviteit en stabiliteit

De toetsen in het Cito Volgsysteem primair onderwijs moeten veranderingen kunnen meten. Uit het kalibratieonderzoek is gebleken dat de opgaven op één onderliggende schaal Woordenschat liggen. De resultaten uit het normeringsonderzoek laten zien dat er verandering gemeten wordt. De gemiddelden per afnamemoment verschillen immers. Uit de (latente) correlaties in tabel 6.4 blijkt dat de correlaties hoog genoeg zijn om te kunnen beweren dat bijna alle leerlingen een zekere groei doormaken. Die groei is niet zo hoog om te kunnen stellen dat dit voor alle leerlingen het geval is, dan wel dat de groei voor alle leerlingen even groot is. De correlaties waarover we beschikken voor aanpalende meetmomenten zijn in alle gevallen minimaal 0,88. In tabel 6.3 vinden we de aantallen leerlingen die op verschillende normeringstijdstippen aan het onderzoek deelgenomen hebben. Voor de volledigheid is in onderstaande tabellen ook het – in een eerdere verantwoording gerapporteerde – normeringsmoment E7 meegenomen. Het betreft de groepen leerlingen die steeds over twee en soms zelfs drie afnamemomenten zijn onderzocht, zodat de correlaties ‘over de tijd’ berekend konden worden.
Bovenstaande is een onderbouwing voor het gegeven dat de toetsen Woordenschat voldoende responsief zijn om veranderingen te meten. Bovendien weerspiegelen de hoge tot zeer hoge correlaties een grote stabiliteit in de uitkomsten van de toetsen. Daarbij moet men zich realiseren dat het op elk afnamemoment om verschillende toetsen gaat die op dezelfde onderlinge vaardigheidsdimensie zijn geconstrueerd.
De hoge correlaties laten niet alleen zien dat de gemeten vaardigheid (woordenschat) zeer stabiel is, zij impliceren ook dat het goed gelukt is om de toetsen op dezelfde vaardigheidsschaal te situeren, hetgeen

kan worden opgevat als een onderbouwing van de validiteit. De correlaties wijzen ten slotte op hoge test-hertestbetrouwbaarheden per toets. Er is weliswaar geen test-hertestonderzoek uitgevoerd. Maar wanneer dergelijke hoge correlaties tussen verschillende toetsen op dezelfde vaardigheidsdimensie tussen twee meetmomenten worden gevonden, mag men aannemen dat herhaalde afname van dezelfde toets met een korter afname-interval tot zeer hoge intercorrelaties (i.e. test-hertestbetrouwbaarheden) zou leiden.

Tabel 6.3 Aantal gevolgde leerlingen op de verschillende normeringsmomenten

	E7	**B8**	**M8**
E7	2403		
B8	1316	1551	
M8		165	755

Tabel 6.4 Latente correlaties tussen leerlingen op de verschillende normeringsmomenten

	E7	**B8**	**M8**
E7	1,000		
B8	0,881	1,000	
M8		0,901	1,000

6.2.5 Gegevens over itemkenmerken

In deze paragraaf vatten we een aantal gegevens samen die betrekking hebben op de itemparameters. Dat doen we voor de digitale en de papieren toetsen afzonderlijk, zoals weergegeven in tabel 6.5 en 6.6.

Tabel 6.5 Samenvatting itemkenmerken voor de digitale toetsen op de afnamemomenten M7 en E7

	E7			**M7**			**B8**			**M8**		
	P	**R_{it}**	**R_{ir}**	**P**	**R_{it}**	**R_{ir}**	**P**	**R_{it}**	**R_{ir}**	**P**	**R_{it}**	**R_{ir}**
gemiddeld	0,667	0,42	0,392	0,675	0,333	0,296	0,620	0,338	0,328	0,655	0,340	0,329
min	0,346	0,263	0,232	0,235	0,191	0,16	0,239	0,200	0,190	0,273	0,198	0,189
max	0,892	0,577	0,553	0,954	0,535	0,503	0,920	0,475	0,465	0,934	0,485	0,475
R_{it} < .20		0			1			0			0	

De gemiddelde moeilijkheidsgraad van de digitale toetsen ligt op het door de toetsdeskundigen gewenste niveau, namelijk rond 0,65. De gemiddelde moeilijkheidsgraad voldoet daarmee aan het gestelde doel, namelijk een optimaal onderscheidend vermogen bij de groep met een lage of gemiddelde vaardigheid (zie verder hoofdstuk 5 over lokale meetnauwkeurigheid), terwijl de toetsen niet als moeilijk zullen worden ervaren door de doorsnee leerling. De moeilijkheidsgraden van de afzonderlijke opgaven kennen een goede spreiding; er zijn zowel moeilijke als gemakkelijke opgaven in de toetsen opgenomen.
De samenhang tussen item- en totaalscore is zowel in termen van R_{ir} als in termen van R_{it} weergegeven. Eerstgenoemde kengetallen geven een reëlere inschatting van die samenhang, maar er zijn geen normwaarden voor beschikbaar in het COTAN-beoordelingssysteem; voor R_{it} is dat wel het geval.
De gemiddelde R_{it}-waarden zijn over het algemeen te kenschetsen als goed (> 0,30). Bij de toets M7 voldoet slechts één opgave niet aan de minimumeis (< 0,20).

Tabel 6.6 Samenvatting itemkenmerken voor de papieren toets op de afnamemomenten B8 en M8

	B8			M8		
	P	**R_{it}**	**R_{ir}**	**P**	**R_{it}**	**R_{ir}**
gemiddeld	0,685	0,341	0,334	0,717	0,340	0,333
min	0,360	0,199	0,192	0,400	0,197	0,190
max	0,905	0,481	0,474	0,922	0,492	0,485

Voor de papieren toets zijn de resultaten vergelijkbaar. De gemiddelde moeilijkheidsgraad voor de twee afnamemomenten ligt zeer constant op een waarde tussen 0,69 en 0,72 met een goede spreiding van de moeilijkheidsgraden over de opgaven. De R_{it}-waarden laten gemiddelden zien van op of hoger dan 0,34, met uitzondering van één opgave in de toets. Die ligt net onder de minimumeis (< 0,20) en dat is nog geen 1,4% van alle opgaven.

7 Samenvatting

In dit hoofdstuk wordt kort weergegeven wat in de voorafgaande hoofdstukken besproken is.
Nadat we in hoofdstuk 2 de uitgangspunten bij de toetsconstructie en in hoofdstuk 3 de inhoud van de toetsen uitvoerig hebben beschreven, hebben we in hoofdstuk 4 over het normeringsonderzoek gerapporteerd. We hebben daar verantwoord hoe de afnamedesigns voor de normeringsonderzoeken en de onderzoeken papier-digitaal zijn opgezet. Ook hebben we in hoofdstuk 4 aangegeven hoe we te werk zijn gegaan bij de steekproeftrekking. De wijze van steekproeftrekking en de controles achteraf (wat betreft percentage achterstandsleerlingen, spreiding over regio's en mate van verstedelijking) wijzen uit dat er sprake is van lichte afwijkingen van de steekproefverdeling. Maar voor het effect van die afwijkingen is gecontroleerd door weging. Het normeringsonderzoek leverde de resultaten op zoals vermeld in tabel 4.12. Op alle normeringsmomenten is sprake van toetsen met normaal verdeelde vaardigheidsdimensies, die elkaar qua spreiding nauwelijks ontlopen en de verwachte toename in gemiddelde vaardigheid laten zien.
In hoofdstuk 5 rapporteerden we over de betrouwbaarheidscoëfficiënten. De betrouwbaarheidscoëfficiënten (MAcc's) zijn zowel voor de papieren als de digitale toetsen Woordenschat hoog tot zeer hoog, ze variëren van 0,90 tot 0,94. In de figuren 5.1 en 5.2 is af te lezen hoe het is gesteld met de lokale meet-nauwkeurigheid van de toetsen. De lokale meetnauwkeurigheid is het grootst in het gewenste bereik.
Over validiteit rapporteerden we in hoofdstuk 6. De toetsen Woordenschat sluiten nauw aan bij het doel en de inhoud van het woordenschatonderwijs op de basisschool (zie paragraaf 6.1). De wijze waarop de toetsen zijn samengesteld garandeert een goede inhoudsvaliditeit.
In paragraaf 6.2 is uitgebreid ingegaan op de begripsvaliditeit van de toetsen Woordenschat. Een belangrijke indicatie voor de validiteit van de woordenschatopgaven uit de toetsen komt uit het kalibratieonderzoek (hoofdstuk 4). Daaruit is gebleken dat de opgavenverzameling waaruit de woordenschattoetsen zijn samengesteld, beschreven kan worden met OPLM. Dat betekent dat de met de toetsen gemeten vaardigheid te verklaren is door een unidimensionaal concept.
Daarnaast kon worden vastgesteld dat de correlaties tussen de latente vaardigheden op twee opeenvolgende afnamemomenten hoog tot zeer hoog zijn. Dat betekent enerzijds dat de toetsen goede operationalisaties zijn van dezelfde onderlinge vaardigheid (i.e. woordenschat) en dat de stabiliteit van deze vaardigheid hoog is. Anders gezegd, als we weten wat de score van een leerling op een bepaald moment is, kunnen we daaruit goed afleiden wat zijn score op een volgend afnamemoment zal zijn. Ook is aangetoond dat de vaardigheidsscore gemiddeld toeneemt van afnamemoment tot afnamemoment. Dit is te verwachten op basis van de toename in leeftijd en ontwikkeling en de vorderingen in het leerproces.
Een andere belangrijke aanwijzing voor begripsvaliditeit is af te leiden uit de correlaties tussen het onderdeel Woordenschat in de Eindtoets Basisonderwijs en andere onderdelen uit de Eindtoets (Cito, 2010). Uit deze gegevens blijkt dat de scores op het onderdeel Woordenschat sterker samenhangen met scores op de onderdelen die inhoudelijk veel raakvlakken hebben met woordenschat. Zo blijkt de samenhang van Woordenschat met Taal (exclusief woordenschat) hoger dan de samenhang met Rekenen-Wiskunde en Studievaardigheden. Daarnaast blijkt – binnen de taalonderdelen – de samenhang van Woordenschat met semantische taalonderdelen, zoals Begrijpend lezen en Schrijfvaardigheid hoger dan de samenhang met een niet-semantisch taalonderdeel als Spelling. De gegevens over de itemkenmerken (moeilijkheidsgraad en item-totaalcorrelatie) laten tot slot een bevredigend beeld zien.

8 Literatuur

Beek, W. van & Verhallen, M. (2004). *Taal, een zaak van alle vakken.* Bussum: Coutinho.

Berkel, S. van (2007). *Balans van het leesonderwijs halverwege de basisschool 4. Uitkomsten van de vierde peiling in 2005.* PPON-reeks nr. 36. Arnhem: Cito.

Berkel, S. van, Groenen, I. & Hilte, M. (2011). *LOVS Woordenschat groep 7*. Arnhem: Cito.

Berkel, S. van, Groenen, I. & Hilte, M. (2012). *LOVS Woordenschat groep 8*. Arnhem: Cito.

Berkel, S. van & Hilte, M. (2009). *LOVS Woordenschat groep 5*. Arnhem: Cito.

Biemiller, A. (2010). *Words Worth Teaching.* New York, NY: McGraw-Hill.

Cito (2010). *Eindtoets Basisonderwijs*. Arnhem: Cito.

Cito (2010). *Entreetoets groep* 7. Arnhem: Cito.

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. & 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.

Filipiak, P. (2004). Wij leren woordjes alleen voor ons proefwerk. Woordenschatonderwijs. *S&B, Vaktijdschrift voor onderwijsadviseurs, 1,* 1.

Filipiak, P. (2006). Woordenschatonderwijs. Beter luisteren, lezen, spreken en schrijven. *JSW, 90 (4),* 15-19.

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.

Haest, I. & Vermeer, A. (2005). Brede en diepe woordkennis, vaktaal en tekstbegrip. *Toegepaste taalwetenschap in Artikelen, 74,* 45-58.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item response Theory*. Newbury Park, CA: Sage.

Hilte, M., Berkel, S. van & Groenen, I. (2010). Het toetsen van woordenschat onder de loep. *Taal lezen primair*, *2,* 16-18.

Huizenga, H. (2005). *Taal en didactiek. Woordenschat*. Groningen: Wolters-Noordhoff.

Kuiken, F. & Droge, S. (2010). *Woordenlijst Amsterdamse Kinderen. Digiwak.* Amsterdam: Universiteit van Amsterdam.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Nulft, D. van den & Verhallen, M. (2002). *Met woorden in de weer. Woordenschatuitbreiding en cognitieve ontwikkeling van leerlingen.* Bussum: Coutinho.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Read, J. (2000). *Assessing Vocabulary.* Cambridge Language Assessment Series. Cambridge: Cambridge University Press.

Rhoer, F. & Vermeer, A. (2005). Woordenschatonderwijs en hiërarchische relaties. *Toegepaste Taalwetenschap in Artikelen, 73,* 99-110.

Schoonen, R. & Verhallen, M. (1998). Kennis van woorden: de toetsing van diepe woordkennis. *Pedagogische Studiën, 75, 3,* 153-168.

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing, 25,* 211-236.

Schrooten, W. & Vermeer, A. (1994). *Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen.* Tilburg: University Press.

Staphorsius, G. (1990). *Pakket Toetsen Woordkennis. Handleiding en verantwoording*. Arnhem: Cito.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid: de ontwikkeling van een domeingericht meetinstrument.* Enschede: Universiteit Twente.

Staphorsius, G., Krom, R.S.H., Kleintjes, F.G.M. & Verhelst, N.D. (1998). *Toetsen Begrijpend Lezen, handleiding*. Arnhem: Cito.

Verhallen, M. (2006). Diepe woordkennis in het basisonderwijs. *Stichting* NOB, lesbrief 1.

Verhallen, M. & Verhallen, S. (1994). *Woorden leren woorden onderwijzen.* Hoevelaken: CPS.

Verhallen, M. & Walst, R. (2001). *Taalontwikkeling op school. Handboek voor interactief taalonderwijs.* Bussum: Coutinho.

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., & 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.

Verhoeven, L. (1996). *Woordenschattoets. Handleiding.* Arnhem: Cito.

Verhoeven, L. & Vermeer, A. (1999). *Leeswoordenschat*. *Handleiding*. Arnhem: Cito.

Vermeer, A. (1997). Breedte en diepte van woordenschat in relatie tot toenemende taalverwerving en frequentie van aanbod. *Gramma/TTT*, *6* (3), 169-187.

Vermeer, A. (2005). Convergente en divergente validiteit van deeltaalvaardigheden in directe en indirecte toetsprocedures. *Gramma/TTT.*

Vermeer, A. & Cohen de Lara, H. (2004). Taaltoetsing getoetst. De toets, de boodschap en het (de)motiveren van leerkrachten. Jong geleerd is oud gedaan. *Talen leren in het basisonderwijs.* (pp. 71-78). Den Haag: Europees Platform voor het Nederlandse Onderwijs.

Vernooy, K. (2007). *Een goede woordenschat: de basis voor een goede schoolloopbaan*. Amersfoort: CPS.

Verstralen, H.H.F.M. (1997). *OPTAL: Inverse OPLAT and item and test characteristics in populations.* Arnhem: Cito.

Bijlagen

Bijlage 1a Voorbeelden van opgaventypen uit de categorie ‘Betekenis’ (uit de toets B8/M8)

Voorbeeld 1: dezelfde betekenis

Wat is een ander woord voor **vermijden**?

A ontbreken
B ontkennen
C ontvoeren
D ontwijken

Voorbeeld 2: definities

Wat is **verdacht**?

A iets wat onbekend is
B iets wat onbetrouwbaar is
C iets wat onverstandig is
D iets wat onverwacht is

Voorbeeld 3: beschrijvingen

Wat betekent **bij iemand in het krijt staan**?

A iemand gelijk geven
B iemand iets schuldig zijn
C iemand terechtwijzen
D iemand voor gek verklaren

Voorbeeld 4: beschrijvingen

Welk woord past het best op de open plaats?
Als je snel op iets reageert, dan ben je …

A alert.
B bedeesd.
C direct.
D oprecht.

Voorbeeld 5: belangrijke betekeniskenmerken

Waar gaat het in de betekenis van **fatsoen** vooral om?

Om:
A hoe iemand met volwassenen moet omgaan.
B hoe iemand zich moet gedragen in de omgang met anderen.
C hoe iemand zich moet kleden om indruk te maken op anderen.
D hoe iemand zijn schoolwerk netjes moet verzorgen.

Voorbeeld 6: belangrijke betekeniskenmerken

Welke woorden zeggen het best iets over de betekenis van **incident**?

A gebeurtenis – onverwacht – plaatsvinden
B kennis – nuttig – verwerven
C uitstapje – jaarlijks – plezierig
D wedstrijd – langdurig – spannend

Bijlage 1b Voorbeelden van opgaventypen uit de categorie ‘Betekenisrelaties’ (uit de toets B8/M8)

Voorbeeld 1: tegenstellingen

Wat is het tegengestelde van **minachting**?

A arrogantie
B invloed
C overmoed
D respect

Voorbeeld 2: betekenisveld

Wat past het best bij de betekenis van **pulver**?

A fijngemalen
B opengesperd
C samengevoegd
D versnipperd

Voorbeeld 3: gezamenlijke kenmerken / generalisatie

Welk woord hoort er qua betekenis niet bij?

A helm
B overall
C toga
D uniform

Voorbeeld 4: gezamenlijke kenmerken / categorisatie

Welk woord hoort niet bij de betekenis van **zich koest houden**?

A beheerst
B gedeisd
C uitgelaten
D zwijgzaam

Voorbeeld 5: gezamenlijke kenmerken / betekeniscluster

Vul het rijtje aan.

uiteraard, natuurlijk, allicht, …, …

A ongetwijfeld, vermoedelijk
B vanzelfsprekend, ongetwijfeld
C vermoedelijk, wellicht
D wellicht, vanzelfsprekend

Voorbeeld 6: Vergrotende trap

Waar staan de woorden in de goede volgorde?

steeds **luider**

A glimlachen – gniffelen – schateren
B gniffelen – glimlachen – schateren
C gniffelen – schateren – glimlachen
D schateren – gniffelen – glimlachen

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

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

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

Fotografie: Ron Steemers