

Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 4

Anke Jolink, Marieke Tomesen, Maartje Hilde, Anke Weekers en
Ron Engelen



Wetenschappelijke verantwoording

Begrijpend lezen 3.0 voor groep 4

Anke Jolink
Marieke Tomesen
Maartje Hilde
Anke Weekers
Ron Engelen

© Cito B.V. Arnhem (2015)

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

Inhoud

1	Inleiding	5
2	Uitgangspunten van de toetsconstructie	7
2.1	Meetpretentie	7
2.2	Doelgroep	7
2.3	Gebruiksdoel en functie	8
2.4	Theoretische inkadering	11
2.4.1	Inhoudelijk	11
2.4.1.1	Leesvaardigheid	11
2.4.1.2	Teksten	13
2.4.1.3	Leesdoel	13
2.4.1.4	Ontwikkeling van de vaardigheid	13
2.4.1.5	Onderwijsdoelen	14
2.4.2	Psychometrisch	16
2.4.2.1	Opgavenbanken	16
2.4.2.2	Het gehanteerde meetmodel	18
3	Beschrijving van de toets	23
3.1	Opbouw en structuur van de toets	23
3.2	Inhoudsverantwoording	25
3.2.1	Uitwerking domeinbeschrijving in vaardigheden, tekstsoorten en opgavenvormen	25
3.2.2	Itemconstructie, onderzoeken en selectie van opgaven voor de toetsen Begrijpend lezen	28
3.2.3	Gerealiseerde verdeling van toetsitems	31
3.3	Statistische beschrijving	33
4	Kalibratie en normering	35
4.1	Opzet voor de normeringsonderzoeken van het LVS: het macrodesign	35
4.2	De kalibratie	37
4.2.1	De opzet van de kalibratie	37
4.2.2	De stappen in de kalibratie	39
4.2.3	Toetsing van het IRT-model	40
4.3	De normering	43
4.3.1	Opzet	44
4.3.2	Representativiteit	48
4.3.3	Normeringsresultaten	50
4.3.4	Geldigheid van de normen	52
5	Betrouwbaarheid en meetnauwkeurigheid	55
5.1	Betrouwbaarheid	55
5.2	Nauwkeurigheid	56
6	Validiteit	63
6.1	Inhoudsvaliditeit	63
6.2	Unidimensionaliteit, respectievelijk structuur	63
6.3	Itemkwaliteit	64
6.4	Itembias	65
6.5	Soortgenootonderzoek	65
6.6	Verschillen tussen relevante subgroepen	67

7 Samenvatting 69

8 Literatuur 71

Bijlagen 75

Bijlage 1 Uitwerking van referentieniveaus 1F en 2F voor Leesvaardigheid 76

Bijlage 2 Moeilijkheid van de opgaven 79

Bijlage 3 Klassieke en IRT-indices van de opgaven in de E3-toets 83

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de toetsen Begrijpend lezen voor groep 4. De toetsen Begrijpend lezen maken deel uit van de derde generatie toetsen van het Cito Volgsysteem primair en speciaal onderwijs en zijn bestemd voor leerlingen in de groepen 3 t/m 8 in het primair onderwijs. Het betreft papieren toetsen¹ voor alle leerjaren.

Tegelijk met deze wetenschappelijke verantwoording voor de toetsen Begrijpend lezen voor groep 4 wordt ook de wetenschappelijke verantwoording voor de toets Begrijpend lezen groep 3 gepubliceerd. Te zijner tijd zullen ook de wetenschappelijke verantwoordingen met de gegevens van de (nog te verschijnen) toetsen Begrijpend lezen voor de groepen 5 t/m 8 gefaseerd worden uitgebracht.

Deze verantwoording biedt tezamen met de inhoud van het toetspakket Begrijpend lezen voor groep 4 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 Begrijpend lezen 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 en speciaal onderwijs niet bedoeld zijn voor 'voorspellend gebruik' is criteriumvaliditeit 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 toetsen Begrijpend lezen voor groep 4. De kwaliteit van het toetsmateriaal en de handleiding is te bepalen door kennis te nemen van de inhoud van het toetspakket.

¹ Binnen het Cito Volgsysteem primair en speciaal onderwijs zullen er geen digitale toetsen voor Begrijpend lezen derde generatie worden uitgebracht.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

Binnen het leesonderwijs op de basisschool wordt een onderscheid gemaakt tussen technisch lezen en begrijpend lezen. Het technisch lezen is geen doel op zich, maar wordt gezien als een voorwaardelijke activiteit voor het leren begrijpen van teksten. Het begrijpen van geschreven teksten is een vaardigheid die traditioneel wordt gemeten met leesbegripstoetsen. Ook de toetsen Begrijpend lezen groep 4 van het Cito Volgsysteem primair en speciaal onderwijs beogen die vaardigheid te meten en de opgaven in deze toetsen zijn er dan ook operationalisering van. De toetsen Begrijpend lezen zijn bedoeld om vast te stellen hoe goed een leerling geschreven teksten kan begrijpen en hoe de vaardigheid in begrijpend lezen van de leerling zich in de loop van de jaren ontwikkelt. (Zie verder paragraaf 2.4.1).

2.2 Doelgroep

De toetsen Begrijpend lezen 3.0 voor groep 4 van het Cito Volgsysteem primair en speciaal onderwijs zijn bestemd voor leerlingen in groep 4 van het basisonderwijs. De toetsen zijn ook geschikt voor leerlingen in het speciaal basisonderwijs en het speciaal onderwijs cluster 2 en 4. Voor deze groepen speciale leerlingen zijn geen afzonderlijke normen vastgesteld. De toetsresultaten van deze leerlingen worden geïnterpreteerd met behulp van de gemiddelde vaardigheidsscores voor leerlingen uit het reguliere onderwijs. Voor deze leerlingen gelden namelijk dezelfde kerndoelen als voor leerlingen in het basisonderwijs, onder voorwaarde dat leerlingen in het speciaal (basis)onderwijs meer tijd krijgen om de kerndoelen te bereiken. Deze leerlingen kunnen én moeten dus langs dezelfde meetlat gehouden worden als de 'reguliere' leerlingen. De leerlingen in het regulier basisonderwijs waarop de normering gebaseerd is, vormen daarmee ook voor de leerlingen in het speciaal basisonderwijs een correcte referentiegroep.

Voor de toetsen van groep 4 zijn zowel voor 'midden leerjaar' (half januari/half februari) als voor 'einde leerjaar' (juni) populatieparameters bepaald. De toetsen kunnen desgewenst ook op een ander moment in het schooljaar worden afgenomen, maar dat maakt het moeilijker om uitspraken te doen over het niveau van een leerling ten opzichte van andere leerlingen in Nederland.

De toetsen Begrijpend lezen groep 4 kunnen ook gebruikt worden voor leerlingen in andere leerjaren die werken op het niveau van groep 4. In de handleiding is toegelicht hoe dit toetsen op maat, met behulp van vaardigheidsscores, in zijn werk gaat. Voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften zijn in de handleiding extra aanwijzingen opgenomen. Voor deze leerlingen zijn alternatieve rapportageformulieren ontwikkeld en zijn extra toetsen beschikbaar. Door het toevoegen van deze extra toetsen beslaan opeenvolgende toetsen kleinere leerstappen. Zo is er naast een toets voor het functioneringsniveau M4 (medio groep 4) en E4 (eind groep 4) ook een toets voor het functioneringsniveau M4E4. Dit is de makkelijke variant van de toets E4. Aan een leerling die zich minder snel ontwikkelt in leesvaardigheid, kan aan het eind van groep 4 dus de toets M4E4 voorgelegd worden. Deze leerling hoeft zo niet een te moeilijke toets (E4) te maken, maar ook niet twee keer dezelfde toets (M4). Voor de extra toetsen zijn de parameters van de reguliere afnamemomenten gebruikt. Het uitgangspunt is dat de normering van de extra toets gebaseerd is op de populatieparameters van een hoger, regulier afname-moment omdat een extra toets een gemakkelijker variant is van de opvolgende reguliere toets. Voor de normering van de extra toets E3M4 zijn de populatieparameters van afnamemoment medio groep 4 gebruikt en voor de extra toets M4E4 de populatieparameters van afnamemoment einde groep 4.

Om de toetsen Begrijpend lezen te kunnen afnemen, is het van belang dat het technisch leesniveau van de leerlingen hoog genoeg is: leerlingen bij wie het lezen nog niet geautomatiseerd is, zijn nog niet toetsbaar voor begrijpend lezen. Ook voor leerlingen die nog maar pas in Nederland verblijven, zijn de toetsen

ongeschikt: leerlingen moeten het Nederlands voldoende beheersen om de opgaven te kunnen maken, voordat de toetsen Begrijpend lezen bij hen worden afgenomen.

Er is bewust voor gekozen om geen gesproken versie van de toetsen Begrijpend lezen uit te brengen. Het doel van de toetsen is immers om vast te stellen hoe goed kinderen geschreven tekst kunnen begrijpen. Dyslectische leerlingen kunnen eventueel de toetsen Begrijpend luisteren maken om de begripsvaardigheid te meten. Tussen de toetsen Begrijpend lezen en Begrijpend luisteren zijn veel overeenkomsten, vooral als het gaat om vaardigheden zoals het afleiden en integreren van informatie uit een tekst. In beide toetsen komen bijvoorbeeld het kunnen benoemen van de hoofdgedachte of de bedoeling van de tekst aan de orde.

De toetsen kunnen worden afgenomen door de leerkracht of IB'er. We gaan daarbij uit van de professionaliteit van de leerkracht/IB'er. Deze wordt geacht in staat te zijn om aan de hand van de aanwijzingen in de handleiding een gestandaardiseerde en ongestoorde toetsafname te realiseren.

2.3 Gebruiksdoel en functie

De toetsen Begrijpend lezen van het Cito Volgsysteem primair en speciaal onderwijs hebben twee doelen: niveaubepaling en progressiebepaling.

Niveaubepaling

De toetsafnamen in het kader van Begrijpend lezen geven de leerkracht informatie over het leesvaardigheidsniveau van zijn leerlingen, individueel en als groep. Iedere behaalde leesvaardigheidsscore kan daartoe normgericht geïnterpreteerd worden op basis van de vaardigheidsverdeling in een referentiegroep (zie paragraaf 4.2 voor de beschrijving van de referentiegroep). De referentiegroep is op basis van de scores van de leerlingen in deze groep op twee manieren in vijf niveaugroepen verdeeld. De eerste manier, met de niveaugroepen I tot en met V, gaat uit van vijf groepen van ieder 20%. Bij de indeling in I tot en met V worden op de registratieoverzichten de laagste groep en de hoogste groep nog onderverdeeld in twee groepen die ieder 10% leerlingen bevatten. Deze groepen worden van elkaar gescheiden door een stippellijn. De tweede indeling levert de niveaugroepen A tot en met E op en is gebaseerd op een indeling in kwartielen. De niveaugroepen A, B en C bestrijken elk een kwart van de populatie. Het vierde kwartiel wordt opgesplitst in twee subgroepen: D (15%) en E (10%). Zie figuur 1 voor een beschrijving van de niveaugroepen.

Eerstgenoemde indeling is dus symmetrisch opgebouwd en heeft als voordeel – boven de indeling gebaseerd op kwartielen – dat er een gemiddelde² groep onderscheiden wordt, namelijk niveaugroep III. Deze indeling blijkt in de praktijk intuïtiever aan te voelen en minder gevoelig te zijn voor verkeerde interpretaties. Om die reden wordt in de handleiding steeds eerst deze indeling genoemd in plaats van de indeling A tot en met E.

² Het betreft hier geen gemiddelde in de statistische betekenis van het woord.

Figuur 1 Onderscheiden niveaugroepen

Niveau	%	Interpretatie
I	20	Ver boven het gemiddelde
II	20	Boven het gemiddelde
III	20	De gemiddelde groep leerlingen
IV	20	Onder het gemiddelde
V	20	Ver onder het gemiddelde

Niveau	%	Interpretatie
A	25	De 25% hoogst scorende leerlingen
1 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

Progressiebepaling

Het volgen van leerlingen in hun groei, ook wel aangeduid als progressiebepaling, is een van de belangrijkste functies van het Cito Volgsysteem primair en speciaal onderwijs (LVS). De toetsen van het LVS geven de leerkracht (en ouders en leerlingen zelf) informatie over de ontwikkeling van de vaardigheden van de leerlingen, individueel en als groep, gedurende (vrijwel) de gehele basisschoolperiode. 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?

Om leerlingen te kunnen volgen wordt de betreffende vaardigheid, in dit geval begrijpend lezen, opgevat als een unidimensionele vaardigheid, of 'latente trek'. 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 (de ruwe score is het aantal opgaven goed) zijn daartoe te transformeren in scores op één vaardigheidsschaal. Deze unidimensionele vaardigheidsschaal die aan de toetsen Begrijpend lezen ten grondslag ligt, is ontwikkeld met behulp van het *One Parameter Logistic Model* (Verhelst, 1993; Verhelst & Glas, 1995; Verhelst, Glas & Verstralen, 1995). Het aantal afnamemomenten per jaar (en het aantal daartoe te construeren verschillende toetsen) wordt bepaald door het tempo waarin een vaardigheid gemiddeld gesproken binnen een leerjaar en over de gehele schoolperiode toeneemt. Meestal is er sprake van twee afnamemomenten per leerjaar ('medio' en 'einde' leerjaar, aangeduid als M en E) en twee - bij het betreffende afnamemoment passende - toetsen. Elke toets worden geconstrueerd op basis van een gekalibreerde itembank, waarbij een toets zo wordt samengesteld dat deze naar inhoud en moeilijkheidsgraad optimaal past bij het afnamemoment waarvoor deze bedoeld is.

Hoe kunnen we de LVS-toetsen Begrijpend lezen 3.0 inzetten om leerlingen te volgen in de tijd? Globaal kunnen we de toetsresultaten van leerlingen (of groepen leerlingen) op twee manieren interpreteren:

- a. We kunnen het toetsresultaat van een leerling vergelijken met die van andere leerlingen op hetzelfde meettijdstip (afnamemoment).
- b. We kunnen de toetsresultaten van dezelfde leerling vergelijken met diens eigen toetsresultaten op eerdere of latere meettijdstippen (afnamemomenten).

Bij beide vergelijkingen maken we gebruik van het feit dat de toetsresultaten door toepassing van het IRT-model (OPLM) in de vorm van vaardigheidsscores afgebeeld kunnen worden op dezelfde vaardigheidsschaal. Dat geldt voor zowel individuele leerlingen als voor (gemiddelde) groeps-resultaten. Dit alles wordt in meer detail uitgelegd in de leerkrachthandleiding (zie het hoofdstuk 'Interpreteren en analyseren op leerling- en groepsniveau').

Ad a.

Bij de eerstgenoemde vergelijking worden de prestaties van een leerling vergeleken met de prestaties van de hele populatie op een gegeven afnamemoment. Hoe doet een leerling het, bijvoorbeeld, ten opzichte van de gemiddelde leerling? Voor dit doel is de populatie, op basis van de data die verzameld zijn in het kader van het normeringsonderzoek (zie hiervoor hoofdstuk 4 van deze wetenschappelijke verantwoording), ingedeeld in vaardigheidsniveaus (I-V, A-E). Vaardigheidsniveau I, bijvoorbeeld, bevat de 20% hoogst scorende leerlingen. Door de vaardigheidsscore van een leerling te vergelijken met deze vaardigheidsniveaus (die zijn afgebakend door percentiepunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk zoals "Dante heeft op afnamemoment medio groep 4 vaardigheidsniveau V behaald". Voor de leerkracht (en voor Dante en haar ouders) bevat deze uitspraak waardevolle informatie. De leerkracht kan op basis hiervan bijvoorbeeld besluiten om Dante extra lesstof aan te bieden.

Ad b.

Voor het vergelijken ('volgen') van een leerling op twee verschillende tijdstippen komen twee methodes in aanmerking. Bij de eerste methode worden de **vaardigheidsniveaus** op de twee tijdstippen vergeleken: "op tijdstip E3 had Dante vaardigheidsniveau IV en op tijdstip M4 was het vaardigheidsniveau V". Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 106, bijvoorbeeld, op tijdstip E3 en vaardigheidsscore 112 op tijdstip M4. Ook hier geldt, net als bij het vergelijken van prestaties met die van andere leerlingen, dat bij eventuele verdere acties van de leerkracht ook andere aspecten moeten worden betrokken.

Bij alle vergelijkingen die mogelijk zijn, zowel die ad a. als die ad b., dienen uitspraken over leerlingen te worden gerelativeerd. In de handleiding is meer informatie te vinden over de wijze waarop de gebruiker dit kan doen. Hieronder gaan we vooral in op het belang van de (on)betrouwbaarheid van de afgenomen toetsen hierbij.

Voor elke toets geldt dat de vaardigheidsscore die bij een toetsresultaat van een leerling hoort, behept is met een meetfout. Als we rekening houden met die meetfout dan zou het best zo kunnen zijn dat Dante vaardigheidsniveau III heeft behaald op het eerste tijdstip en niet vaardigheidsniveau IV. Of dat we moeten concluderen dat het verschil in de prestaties op de twee tijdstippen statistisch niet significant is: Dante is voor- noch achteruitgegaan. In alle gevallen speelt het betrouwbaarheidsinterval (BI) rondom de vaardigheidsscore een belangrijke rol. Dat geldt ook voor de indeling in vaardigheidsniveaus. Hoe het BI doorwerkt in de indeling in vaardigheidsniveaus en de verdere gevolgen daarvan wordt beschreven in hoofdstuk 5.

Op deze plaats beperken we ons tot de vraag of we de uitspraak kunnen doen dat een leerling of groep (werkelijk) 'gegroeid' is. De eenvoudigste manier is om te kijken of de BI's voor de twee tijdstippen overlappen. Als deze twee BI's niet overlappen dan is er sprake van een significant verschil in vaardigheid tussen beide tijdstippen. Overlappen ze wel dan is er geen verschil in vaardigheid. Deze eenvoudige manier van vergelijken kan de leerkracht zelf uitvoeren en wordt ook in de handleiding beschreven.

We geven hier een voorbeeld. Bij de afname Begrijpend lezen E3 behaalde Harun een vaardigheidsscore

van 112 met een 67% betrouwbaarheidsinterval van 103-120. Bij de afname M4 behaalde Harun een vaardigheidsscore van 131; het bijbehorende betrouwbaarheidsinterval daarbij is 121-141. Aangezien de betrouwbaarheidsintervallen niet overlappen kunnen we zeggen dat Haruns vaardigheid is toegenomen.

Conclusie

De vaardigheidsgroei voor begrijpend lezen voltrekt zich langzaam in de tijd. De verschillen tussen vaardigheidsscores op achtereenvolgende meettijdstippen zijn betrekkelijk klein, al is de gemiddelde vaardigheidstoename in groep 3 en groep 4 wat groter dan in de hogere leerjaren. Bovendien is er sprake van meetfouten. De verschillen in vaardigheidsgroei moeten tegen de achtergrond van die meetfouten worden geïnterpreteerd. Dit betekent dat men weliswaar uitspraken kan doen over de vaardigheidsgroei van een leerling, maar dat deze uitspraken met voorzichtigheid dienen te worden gehanteerd. Dit geldt ook wanneer men de progressie van een leerling volgt in termen van vaardigheidsniveaus of een vergelijking maakt met andere leerlingen in termen van vaardigheids-niveaus. Want ook bij de indeling in vaardigheidsniveaus speelt de nauwkeurigheid van de toets een rol. Hoe de leerkracht hier in de praktijk mee om moet gaan wordt toegelicht in de handleiding voor de leerkracht.

2.4 Theoretische inkadering

2.4.1 Inhoudelijk

Leesvaardigheid wordt in het algemeen omschreven als de vaardigheid om schriftelijke teksten te begrijpen en te gebruiken in overeenstemming met het leesdoel (Campbell et al. 2001; Mullis et al., 2006). Bij het leesproces is er sprake van een constante interactie tussen de lezer met zijn *vaardigheden*, de *tekst* en het *leesdoel* (Rapp et al., 2007). Deze drie componenten worden in onderstaande paragrafen verder toegelicht. Daarnaast bespreken we de ontwikkeling van de vaardigheid en de onderwijsdoelen en leerstoflijnen die de basis vormen voor de toetsmatrijs van de toetsen Begrijpend lezen van het Cito Volgstelsel primair en speciaal onderwijs.

2.4.1.1 Leesvaardigheid

Lezen is een complex en constructief proces waarbij de lezer in interactie met de tekst betekenis toekent aan de tekst (Aarnoutse & Verhoeven, 2003; Campbell et al., 2001; Mullis et al., 2006). De lezer heeft een actieve en initiërende rol. Hij construeert betekenis op basis van de informatie in de tekst én op basis van zijn eigen kennis. Er is hierbij sprake van een continue wisselwerking tussen verschillende verwerkingsprocessen: de verwerking van informatie die is weergegeven in de tekst, de inzet van eigen kennis van de wereld en, uiteindelijk, het integreren van informatie uit de tekst in de eigen kennis. Wanneer deze processen succesvol verlopen, vormt de lezer een samenhangende mentale representatie van de tekst (Kintsch, 2004).

Voor het begrip van een tekst vertrouwen lezers in eerste instantie op de automatische processen die in werking treden tijdens het lezen, maar wanneer dit niet leidt tot voldoende begrip moeten zij meer bewust sturing geven aan het proces van betekenisconstructie (Kintsch, 2004; Rapp et al., 2007). Goede lezers maken in die gevallen gebruik van leesstrategieën, zoals het samenvatten van informatie in de tekst en de inzet van voorkennis over een onderwerp (Pressley, 2000; Van den Broek, 2012; Van den Broek & Espin, 2012).

De cognitieve en metacognitieve aspecten van leesvaardigheid zijn echter niet de enige factoren die van invloed zijn op het proces van betekenisconstructie: ook sociaal-culturele factoren (Alexander & Jetton, 2000; Van Diepen, 2007) en motivationele factoren spelen een belangrijke rol bij begrijpend lezen (Elsäcker, 2002; Van Diepen, 2007; Verhoeven & Snow, 2001). Er is dan ook geen sprake van één proces dat zich steeds op dezelfde wijze voltrekt. De achtergrond, motivatie, eigen kennis en vaardigheden van lezers maken dat het leesproces, tot op zekere hoogte, voor iedere lezer en in iedere situatie anders verloopt.

Uit onderzoek naar het proces van begrijpend lezen is gebleken dat goede lezers zich op een aantal punten onderscheiden van minder goede lezers (Boerma, Manders & Mankhorst, 2009; Van den Broek & Espin, 2012). De hieronder genoemde factoren blijken van invloed te zijn op het leesbegrip:

- de technische leesvaardigheid;
- de breedte en diepte van de woordenschat;
- de algemene begripsvaardigheid: het vermogen om informatie af te leiden uit de tekst en deze te verbinden aan de eigen kennis om tot een coherente representatie van de tekst te komen;
- het vermogen om zelf sturing te geven aan het leesproces door het proces te monitoren en leesstrategieën toe te passen;
- kennis van taal en van tekstkenmerken en tekststructuren;
- kennis van de wereld en, meer specifiek, eigen kennis van het onderwerp van een tekst;
- interesse in het onderwerp van een tekst;
- de motivatie om te lezen.

Leesvaardigheid kan worden onderverdeeld in de vaardigheden *begrijpen*, *interpreteren* en *reflecteren* (zie ook Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009b). Deze vaardigheden staan niet op zichzelf, maar liggen in elkaars verlengde en worden in wisselwerking met elkaar toegepast tijdens het lezen van een tekst.

Begrijpen

De vaardigheid *begrijpen* heeft betrekking op de verwerking van informatie die expliciet in de tekst vermeld staat. Om tot begrip van de tekst te komen, maakt de lezer gebruik van de inhoud (de betekenis van woorden, woordgroepen, zinnen, alinea's en hun onderlinge betekenisrelaties), van expliciete relaties tussen tekstelementen (woord- en zinsvolgorde, verwijzingen, en talige en grafische structuurmarkeerders) en van de expliciete structuur van een tekst.

Interpreteren

Van werkelijk en diepgaand tekstbegrip is pas sprake wanneer informatie uit de tekst verbonden wordt aan de 'eigen' kennis van de lezer. Een lezer benadert een tekst niet blanco, maar zet bij de verwerking van de tekst ook zijn eigen kennis in, waaronder zijn kennis van de wereld en kennis van taal en tekstkenmerken. Dit samenspel van verwerkingsprocessen is aan de orde bij de vaardigheid *interpreteren*.

Impliciete informatie speelt een belangrijke rol bij het lezen van teksten. De schrijver veronderstelt bepaalde kennis bij de lezer bekend en zal die kennis niet altijd expliciet maken. Het is vervolgens aan de lezer om zich te realiseren welke kennis bekend wordt verondersteld, deze informatie te activeren en, indien nodig, aan te vullen met eigen kennis. Het onderkennen en afleiden van impliciete informatie in een tekst, het maken van inferenties, is een belangrijk aspect van deze vaardigheid.

Reflecteren

Het kenmerkende van de vaardigheid *reflecteren* is de beschouwende, evaluerende en kritische kijk op teksten. De lezer neemt dan als het ware afstand van de tekst, vormt zich er een mening over en/of toetst die aan een bepaald standpunt. Hij beschouwt en evalueert betekenis en inhoud, taal, tekstuele en contextuele elementen en het belang, de kwaliteit en de integriteit van de tekst. Het gaat hier niet meer om begrip als zodanig, maar om denken over, reflecteren en abstract redeneren. Dit kan uitmonden in uitspraken over de tekst in evaluerende en waarderende zin.

Hoewel *reflecteren* een belangrijk onderdeel vormt van begrijpend lezen, wordt deze vaardigheid niet gemeten met de toetsen Begrijpend lezen. Het is namelijk niet goed mogelijk gebleken om het reflecteren op teksten te toetsen met meerkeuzeopgaven. Bij deze vaardigheid gaat het immers om het geven van een argumentatie of mening en om dit te toetsen zouden dan ook open opgaven en beoordelingsschema's gebruikt moeten worden.

2.4.1.2 Teksten

De teksten in de toetsen Begrijpend lezen betreffen schriftelijk materiaal. Dit kunnen teksten uit jeugdboeken, tijdschriften of kranten zijn, maar ook teksten van websites. Het onderscheid in teksten is van belang omdat verschillende teksten verschillende accenten in de leesprocessen uitlokken (Goldman & Rakestraw, 2000) en omdat leerlingen zowel binnen als buiten school met verschillende soorten teksten moeten kunnen omgaan. De teksten waarmee leerlingen in het dagelijks leven worden geconfronteerd, kunnen worden onderverdeeld in twee hoofdcategorieën: zakelijke teksten enerzijds en literaire en fictionele teksten anderzijds (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a). Voor de toetsen Begrijpend lezen zijn binnen deze hoofdcategorieën vijf teksttypen onderscheiden: informatieve teksten, instructieve teksten, betogende teksten, verhalende teksten en poëzieteksten. Om variëteit in teksten te waarborgen, zijn naast de verschillende teksttypen ook verschillende tekstgenres geselecteerd. Genres zijn herkenbare vormen van communicatie die zich onderscheiden in doel, structuur en inhoud. Enkele voorbeelden zijn een nieuwsbericht, een verslag, een uitnodiging, een spelinstructie, een recept, een verhaal en een gedicht.

Bij de samenstelling van de toetsen Begrijpend lezen is ernaar gestreefd om zo veel mogelijk gebruik te maken van bestaande, authentieke teksten. Het onderwerp van de tekst is van grote invloed op het proces van begrijpend lezen: de mate waarin het onderwerp aansluit op de kennis van leerlingen heeft aantoonbare effecten op het tekstbegrip (Fisher, Frey & Lapp, 2009). Voor de toetsen Begrijpend lezen zijn daarom teksten geselecteerd met onderwerpen die aansluiten op de kennis en belevingswereld van leerlingen.

Een apart probleem betreft de mogelijke bekendheid van bestaande teksten. Om dit probleem tegen te gaan worden in ieder geval geen teksten uit methodes gebruikt. Een grote variatie in soorten teksten, inhouden en thema's, genres, en andere zaken zal de mogelijke bekendheid van teksten zoveel mogelijk spreiden over de verschillende leerlingen. De teksten variëren ook ten aanzien van het taalgebruik, de lengte van de tekst, het tijdsperspectief bij fictie, de dichtheid van informatie en mate van abstractie. Daarnaast is er rekening gehouden met overwegingen met betrekking tot geslacht, rol, ras, etniciteit, cultuur en religie: teksten met onderwerpen die alleen aansluiten bij de belevingswereld van specifieke subgroepen binnen de populatie, zijn niet opgenomen in de toets.

2.4.1.3 Leesdoel

De manier waarop een lezer leest, hangt voor een groot deel af van het doel om te lezen en het type tekst (Rapp et al, 2007; Van den Broek et al., 2001). Het leesdoel en het daarmee samenhangende verwachtingspatroon zijn van invloed op de vaardigheden en strategieën die de lezer, voor een deel onbewust, aanspreekt om de betekenis van de tekst te achterhalen. Zo zal een lezer die ter ontspanning een fictietekst leest andere verbanden leggen dan een lezer die een informatieve tekst probeert te begrijpen.

In de context van een toetsafname zal het persoonlijke leesdoel voor leerlingen in eerste instantie bestaan uit het correct beantwoorden van de opgaven bij de teksten. Door de variatie in teksttypen en -genres in de toetsen Begrijpend lezen worden de leesdoelen die gerelateerd zijn aan specifieke tekstsoorten echter zo veel mogelijk gedekt.

2.4.1.4 Ontwikkeling van de vaardigheid

Een goede technische leesvaardigheid wordt vaak genoemd als voorwaarde voor begrijpend lezen: een lezer moet de tekst immers kunnen ontcijferen om deze te kunnen begrijpen. Uit diverse onderzoeken blijkt echter dat de ontwikkeling van de algemene begripsvaardigheden die ten grondslag liggen aan begrijpend lezen, al start voordat kinderen beginnen met lezen (McNamara & Kendeou, 2011; Van den Broek, 2012; Van den Broek & Espin, 2012). Zo kunnen peuters al eenvoudige relaties leggen en blijken kleuters al pogingen te doen om coherente mentale representaties te vormen bij verhalen (Van den Broek, 2012). Hoewel de begripsvaardigheid zich al op jonge leeftijd begint te ontwikkelen, is deze in eerste instantie nog niet vergelijkbaar met die van ervaren lezers: de begripsvorming verloopt anders en minder efficiënt dan bij oudere kinderen en volwassenen.

De meeste Nederlandse scholen starten in groep 3 met technisch lezen en geven specifieke lessen voor begrijpend lezen vanaf jaargroep 4. De ontwikkeling van begrijpend lezen kan omschreven worden als een cyclisch, concentrisch proces: leerlingen doorlopen herhaaldelijk dezelfde ontwikkelings- en leerprocessen, maar op een steeds hoger niveau (Sijstra, Aarnoutse & Verhoeven, 1999, in: Aarnoutse & Verhoeven, 2003). De verschillende aspecten van begrijpend lezen, zoals het bepalen van het onderwerp van een tekst of het leggen van verbanden, worden dan ook in alle jaargroepen aan de orde gesteld. De leerlijnen worden niet zozeer gekenmerkt door een uitbreiding van vaardigheden, maar meer door een verdere ontwikkeling van de vaardigheden.

Bij de invulling van de lessen voor begrijpend lezen wordt in de meeste gevallen gebruik gemaakt van methodes en wordt veel aandacht besteed aan het leren toepassen van leesstrategieën (Förner & Van de Mortel, 2010). Hieronder enkele voorbeelden van strategieën die in de meest gebruikte methodes aan de orde worden gesteld:

Strategieën voorafgaand aan het lezen:

- het leesdoel vaststellen;
- de inhoud van de tekst voorspellen;
- de eigen voorkennis activeren.

Strategieën tijdens het lezen:

- controleren of de tekst nog begrepen wordt en onduidelijkheden ophelderen.

Strategieën na het lezen:

- de tekst samenvatten;
- nagaan of het leesdoel bereikt is.

Deze strategieën worden in de methodes over de leerjaren heen herhaaldelijk aangeboden. De teksten waarop leerlingen de strategieën leren toepassen, worden complexer van vorm en inhoud in de bovenbouw. Zo verschillen de teksten voor de lagere en hogere jaargroepen onder meer in woordkeuze, zins- en tekstlengte, structuur, onderwerpkeuze, genre, mate van abstractie en de hoeveelheid expliciete en impliciete verbanden.

2.4.1.5 Onderwijsdoelen

Het onderwijsaanbod dat scholen bieden, moet voldoen aan de kerndoelen die zijn opgesteld door het ministerie van Onderwijs, Cultuur en Wetenschap (Ministerie van OCW, 2006). Deze kerndoelen beschrijven wat leerlingen aan het einde van het basisonderwijs aan kennis en vaardigheden moeten hebben verworven. Voor begrijpend (en studerend) lezen zijn de volgende kerndoelen geformuleerd:

Schriftelijk taalonderwijs:

- Leerlingen leren informatie te achterhalen in informatieve en instructieve teksten waaronder ook schema's, tabellen en digitale bronnen;
- Leerlingen leren informatie en meningen te ordenen bij het lezen van school- en studieteksten, andere instructieve teksten en bij systematisch geordende bronnen, waaronder ook digitale;
- Leerlingen leren informatie en meningen te vergelijken en te beoordelen in verschillende teksten;
- Leerlingen krijgen plezier in het lezen en schrijven van voor hen bestemde verhalen, gedichten en informatieve teksten.

Taalbeschouwing, waaronder strategieën:

- Leerlingen leren bij de doelen onder 'schriftelijk taalonderwijs' strategieën te herkennen, te verwoorden, te gebruiken en te beoordelen.

In de tussendoelen gevorderde geletterdheid voor de middenbouw (Aarnoutse & Verhoeven, 2003) zijn markeringspunten beschreven die inhoudelijk de verbinding leggen met de kerndoelen Nederlandse taal. Voor begrijpend lezen zijn de volgende tussendoelen bepaald:

Tussendoelen begrijpend lezen:

De leerlingen lezen eenvoudige teksten die verhalend, informatief, directief, beschouwend en argumentatief van aard zijn met begrip en voeren daarbij de volgende leesstrategieën uit:

- Ze bepalen het thema van een tekst en activeren hun eigen kennis over het thema;
- Ze koppelen verwijswaarden aan antecedenten;
- Ze lossen het probleem van een moeilijke zin (of zinnen) op;
- Ze voorspellen de volgende informatie in een tekst;
- Ze leiden informatie af uit een tekst;
- Ze onderscheiden verschillende soorten teksten zoals verhalende, directieve, beschouwende en argumentatieve teksten;
- Ze herkennen de structuur van verhalende teksten.

In 2010 is de wet Referentieniveaus Nederlandse taal en rekenen van kracht geworden, met als doel een betere aansluiting te bewerkstelligen tussen het taal- en rekenonderwijs in de verschillende onderwijssectoren en de taal- en rekenvaardigheden van leerlingen te verbeteren. De wet is gebaseerd op het werk van de Expertgroep Doorlopende Leerlijnen: Over de drempels met Taal en Rekenen, hoofdrapport (2008a), Over de drempels met taal, de niveaus voor de taalvaardigheid (2008b) en Een nadere beschouwing (2009b). De referentieniveaus voor het onderdeel lezen in het basisonderwijs zijn beschreven in het 'Referentiekader taal en rekenen' van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009a). Hierin is vastgelegd wat leerlingen moeten leren als het gaat om Nederlandse taal en rekenen. Voor het leesonderwijs in de basisschool zijn vooral referentieniveaus 1F en 2F van belang. Referentieniveau 1F is een minimum- of drempelniveau dat ongeveer 75% van de leerlingen van de basisschool aan het einde van groep 8 kan halen. Het niveau 2F geldt als streefniveau voor de basisschool en wordt ook wel aangeduid als niveau 1S.

De referentieniveaus omvatten behalve een algemene omschrijving een opsomming van de typen teksten die leerlingen lezen (teksttypen) en de taken die zij uitvoeren (taaktypen). Daarnaast omvatten de referentieniveaus een omschrijving van de kenmerken van de taakuitvoering. Voor het domein lezen is hierbij onderscheid gemaakt tussen het lezen van zakelijke teksten enerzijds, en het lezen van fictionele, narratieve en literaire teksten anderzijds. In bijlage 1 staan de uitwerkingen voor begrijpend lezen beschreven, voor de verschillende niveaus en de verschillende teksttypen.

In de publicatie 'Leerstoflijnen lezen beschreven' van de SLO (Oosterloo & Paus, 2010) is het referentiekader Nederlandse taal uitgewerkt in richtlijnen voor de opbouw van de leerstof voor het domein lezen, over de verschillende jaargroepen binnen het basisonderwijs. De leerstoflijnen zijn achtereenvolgens uitgewerkt voor jaargroepen 3 en 4, 5 en 6 en 7 en 8, aan de hand van de drie overkoepelende kenmerken uit het referentiekader: de leestaken, de kenmerken van teksten en de kenmerken van de taakuitvoering.

Het referentiekader Nederlandse taal, de tussendoelen en de Leerstoflijnen lezen zijn als uitgangspunten gebruikt bij de opzet en ontwikkeling van de toetsen Begrijpend lezen voor groep 4: ze vormen de basis voor de toetsmatrijs (zie paragraaf 3.2).

2.4.2 Psychometrisch

2.4.2.1 Opgavenbanken

Voor het samenstellen van toetsen voor het primair en speciaal onderwijs beschikt Cito over opgavenbanken. Die liggen ten grondslag aan onder meer de toetsen in het Cito Volgsysteem voor primair en speciaal onderwijs (LVS). Voor de constructie van de toetsen LVS Begrijpend lezen is gebruikgemaakt van de opgavenbank Begrijpend lezen. Voor andere vakgebieden in het LVS als Spelling, Woordenschat, Rekenen-Wiskunde en Studievaardigheden zijn eveneens opgavenbanken in gebruik.

Een opgavenbank is nadrukkelijk niet eenvoudigweg een verzameling opgaven of items waaruit een toetsconstructeur min of meer naar willekeur een aantal items selecteert om een nieuwe toets te construeren. In deze paragraaf wordt beschreven 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 begrijpend lezen 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 van leesvaardigheid uit, waarbij een groter getal wijst op een grotere leesvaardigheid. 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 grootheid geeft aan hoe nauwkeurig die schatting is, en heeft dus de status van een standaardfout, te vergelijken met de standaardmeetfout uit de klassieke testtheorie.

Latente vaardigheid

De antwoorden van een leerling op de items worden beschouwd als indicatoren van de vaardigheid, hetgeen ruwweg betekent dat men verwacht dat alle items in de opgavenbank leesbegrip meten. De vaardigheid zelf wordt als niet-observeerbaar beschouwd, en daarom gewoonlijk omschreven als een latente vaardigheid.

'Moeilijkheid' in de Item Respons Theorie

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

Kansmodel

De ruwe omschrijving van de moeilijkheidsgraad die in de vorige alinea werd gehanteerd (de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden) behoeft enige verdere uitwerking. Men zou deze omschrijving kunnen opvatten als een drempel: heeft een leerling die mate van vaardigheid niet, dan kan hij het item 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 eruit volgt dat een leerling die een moeilijk item correct beantwoordt geen fout kan maken op een gemakkelijker item. Daarom wordt in de IRT een

kansmodel gebruikt: hoe groter de vaardigheid, des te groter de kans dat een item juist wordt beantwoord. De moeilijkheidsgraad van een item wordt dan gedefinieerd als de mate van vaardigheid die nodig is om met een kans van precies een half een juist antwoord te kunnen produceren.

Kalibratie

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

Afnamedesigns

Meestal bevat een opgavenbank meer items dan een doorsnee toets, zodat het praktisch niet doenbaar is om alle items aan alle leerlingen voor te leggen. Elke leerling in de kalibratiesteekproef krijgt derhalve slechts een (klein) gedeelte van de items uit de opgavenbank voorgelegd. Dit gedeeltelijk voorleggen gebeurt aan de hand van een zogeheten 'onvolledig design'. Dit 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 de kalibratie met succes uitgevoerd is, is het resultaat een zogenoemde gekalibreerde itembank. In dat proces worden de items die niet passen bij de verzameling uit de collectie verwijderd. De opgavenbank bevat voor elk item niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen, en de statistische zekerheid dat alle items dezelfde vaardigheid aanspreken. Dit houdt onder meer het volgende in:

- 1 In principe kan met een willekeurige selectie items uit de bank de vaardigheid worden gemeten bij een willekeurige leerling. In principe, want een willekeurige toets die uit de itembank wordt getrokken zal in de praktijk meestal niet voldoen omdat de meetresultaten (de schatting van de vaardigheid) onvoldoende nauwkeurig zal zijn. Voor een nauwkeuriger meting (bij een gegeven aantal items in de toets) moeten de moeilijkheidsgraden van de items in overeenstemming gebracht worden met het vaardigheidsniveau van de leerlingen.
- 2 Om een schatting te kunnen maken van de verdeling van de vaardigheid in een welomschreven populatie, worden selecties van items voorgelegd aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. In het geval van Begrijpend lezen LVS zijn dat steekproeven van leerlingen op de verschillende normeringsmomenten vanaf eind groep 3 tot en met medio groep 8. Daarbij maakt het, behoudens wat bij 1 is vermeld over nauwkeurigheid, niet uit welke selectie van items aan een leerling binnen een normeringsgroep wordt afgenomen. Een van de eigenschappen van gekalibreerde itembanken is immers dat met elke selectie items de vaardigheid van leerlingen kan worden bepaald. Voor een voorbeeld hiervan, zie 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 tot de betreffende referentiepopulatie behoren, 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 4 kan een toets maken die normaliter aan groep 3 wordt voorgelegd, en zijn vaardigheidsschatting kan behalve met de populatie van groep 4 ook vergeleken worden met de percentielen in de populatie van groep 3, met bijvoorbeeld de uitspraak: "De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 3".
- 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 groep 4 wordt voorgelegd. Immers, het kalibratieonderzoek heeft aangetoond dat alle items dezelfde vaardigheid meten. Een nieuwe toets meet dus dezelfde vaardigheid, zodat schattingen die van verschillende toetsen afkomstig zijn zinvol met elkaar kunnen worden vergeleken.

Tot zover de nadere bepaling van het begrip 'opgavenbank'. In de volgende hoofdstukken van dit deel van de verantwoording worden de begrippen die hierboven aan de orde zijn geweest nader uitgewerkt en toegelicht voor de opgavenbank Begrijpend lezen. De verantwoording van de inhoudelijke constructie van deze opgavenbank staat in hoofdstuk 3. In hoofdstuk 4 wordt (onder andere) de psychometrische constructie van de opgavenbanken besproken (kalibratie).

2.4.2.2 Het gehanteerde meetmodel

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

In de IRT staat het te meten begrip of de te meten eigenschap centraal. De IRT beschouwt het antwoord op een item als een indicator voor de mate waarin die eigenschap aanwezig is. Het verband tussen eigenschap en itemantwoord is van probabilistische aard en wordt weergegeven in de zogenaamde itemresponsfunctie. Die geeft aan hoe groot de kans is op een correct antwoord als functie van de onderliggende eigenschap of vaardigheid. Formeler: zij X_i de toevalsvariabele die het antwoord op item i voorstelt. X_i neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord. Als symbool voor de vaardigheid wordt θ (theta) gekozen. De vaardigheid θ is niet rechtstreeks observeerbaar. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd³. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

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

³ Dit maakt duidelijk waarom men de modellen die ressorteren onder de IRT, ook wel aanduidt met 'latente trek'-modellen.

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

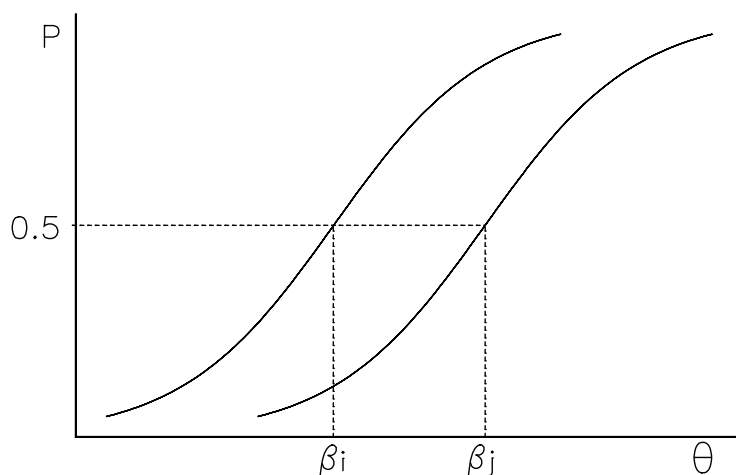
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \quad (2.2)$$

waarin β_i de moeilijkheidsparameter van item i is. Dat is een onbekende grootte die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.1 voor twee items, i en j , die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende functie is van θ : hoe groter de vaardigheid, des te groter de kans op een juist antwoord. Indien de latente vaardigheid precies gelijk is aan de moeilijkheidsparameter β_i , volgt

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

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

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 item j juist te beantwoorden steeds kleiner is dan de kans op een juist antwoord op item i . Hieruit volgt de statistisch te toetsen voorspelling dat de verwachte proportie juiste antwoorden op item j kleiner is dan op item i in een willekeurige steekproef van personen. Splitst men nu een grote steekproef in twee deelsteekproeven, een 'laaggroep', met de vijftig procent laagste scores, en een 'hooggroep', met de vijftig procent hoogste scores, dan kan men nagaan of de geobserveerde p-waarden van de opgaven in beide deelsteekproeven op dezelfde wijze geordend zijn. Daarvan kan strikt genomen alleen sprake zijn als, in termen van de klassieke testtheorie uitgedrukt, alle opgaven eenzelfde discriminatie-index hebben. Dat echter blijkt lang niet altijd zo te zijn. Ook in ons geval niet.

Veel van de items blijken dan ook niet beschreven te kunnen worden met het Raschmodel. Daarom is bij dit instrument gekozen voor een ander IRT-model.

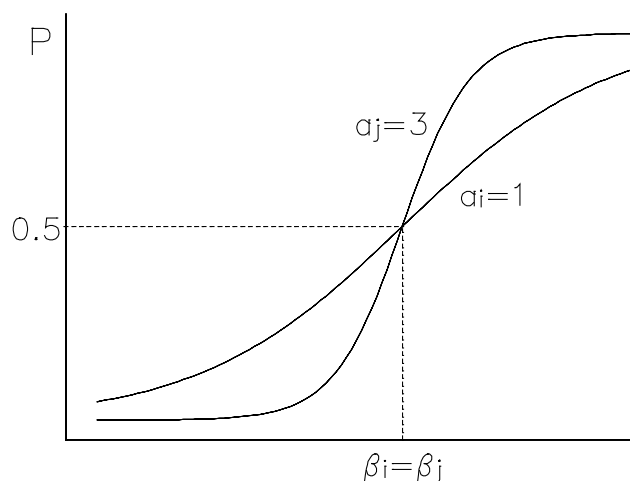
Alvorens het hier gebruikte model te introduceren, is eerst een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattingsmethode is de 'conditionele grootste aannemelijkheidsmethode' (in het Engels: Conditional Maximum Likelihood, verder aangeduid als CML). Die maakt gebruik van het feit dat in het Raschmodel een afdoende steekproefgrootte ('sufficient statistic') bestaat voor de latente variabele θ , namelijk de ruwe score of het aantal correct beantwoorde items. Dat betekent grofweg dat, indien de itemparameters bekend zijn, alle informatie die het antwoordpatroon over de vaardigheid bevat, kan worden samengevat in de ruwe score; het doet er dan verder niet meer toe welke opgaven goed en welke fout zijn gemaakt. Hieruit vloeit voort dat de conditionele kans op een juist antwoord op item i , gegeven de ruwe score, een functie is die alleen afhankelijk is van de itemparameters en onafhankelijk van de waarde van θ^4 . 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 zogenoemde éénparameter logistisch model (One Parameter Logistic Model, afgekort: OPLM) is CML mogelijk. Dit model is, anders dan het Raschmodel, wel bestand tegen 'omwisseling' van 'proporties juist' in verschillende steekproeven (Glas & Verhelst, 1993; Eggen, 1993; Verhelst & Kleintjes, 1993). De itemresponsfunctie van het OPLM is gegeven door:

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

waarin a_i de zogenaamde discriminatie-index van het item is. Door deze indices te beperken tot (positieve) gehele getallen, en door ze a-priori als constanten in te voeren, is het mogelijk CML-schattingen van de itemparameters β_i te maken. In figuur 2.2 is de itemresponscurve weergegeven van twee items i en j , die even moeilijk zijn maar verschillend discrimineren.

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



⁴ Een gedetailleerde uiteenzetting hierover kan men vinden in Verhelst, 1992.

De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas en Verstralen, 1995). Dit programma voert 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 items volgens het OPLM is dan ook een iteratief proces waarin alternerend de modelfit van items wordt onderzocht door middel van statistische toetsing en de waarden van de discriminatie-indices worden aangepast op grond van de resultaten van deze toetsen. Deze aanpassingen geschieden in de praktijk op basis van een en hetzelfde gegevensbestand. Er kan dus kanskapitalisatie optreden. Indien een steekproef een voldoende grootte heeft, is het effect van deze kanskapitalisatie echter gering (Verhelst, Verstralen en Eggen, 1991).

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

Voor de schatting van de populatieverdeling wordt gebruikgemaakt van "plausible values". Plausible values representeren de volledige range aan vaardigheidsscores die een leerling zou kunnen hebben, gegeven zijn itemantwoorden. Een plausible value is dus niet gelijk aan de θ -parameter (de vaardigheidsscore) zoals die gedefinieerd is in bijvoorbeeld het OPLM. Er wordt namelijk niet één enkele puntschatting voor θ bepaald, maar er wordt een reeks van mogelijke waarden voor θ geschat die elk een bepaalde kans hebben om geobserveerd te worden. *Plausible values* zijn random trekkingen uit deze (geschatte) verdeling voor de vaardigheid θ van een leerling. Daardoor geven *plausible values* niet alleen informatie over de geschatte vaardigheid van een leerling, maar ook over de onzekerheid die bij die schatting hoort. Het gebruik van de puntschattingen voor de latente vaardigheid θ zou tot bias in sommige populatieparameters leiden. Zo zou de variantie bij gebruik van de ML- of WML-schatting van θ bijvoorbeeld overschat worden.

3 Beschrijving van de toets

3.1 Opbouw en structuur van de toets

Het toetspakket Begrijpend lezen 3.0 voor groep 4 uit het Cito Volsysteem primair en speciaal onderwijs bevat in totaal vier papieren toetsen: E3M4, M4, M4E4 en E4. In groep 4 zijn er twee afnamemomenten: halverwege het schooljaar (half januari/half februari) en aan het einde van het schooljaar (juni). De toetsen M4 en E4 zijn de reguliere toetsen, bedoeld voor afname op de reguliere afnamemomenten medio (M) of einde (E) schooljaar. Naast deze reguliere toetsen bevat het toetspakket ook extra toetsen (E3M4 en M4E4) voor leerlingen met een vertraagde ontwikkeling. Ook deze extra toetsen zijn bedoeld voor afname op de reguliere afnamemomenten.

Door het toevoegen van een extra toets (tussentoets) beslaan opeenvolgende toetsen kleinere leerstappen. Daardoor zijn er voor speciale leerlingen meer toetsen beschikbaar voor het meten van hun leesvaardigheid. Zo is er een toets voor het functioneringsniveau M4 (medio groep 4) en E4 (eind groep 4), maar ook voor het functioneringsniveau M4E4. De toetsen E3M4 en M4E4 zijn net wat makkelijker dan de reguliere toetsen M4 respectievelijk E4. Zo kan de toets M4E4 voorgelegd worden aan leerlingen voor wie de toets E4 nog net te moeilijk is. Deze leerlingen hoeven dan niet nogmaals de toets M4 te maken.

Opbouw

De toetsen Begrijpend lezen 3.0 voor groep 4 bestaan telkens uit twee taken. Deze dienen bij voorkeur te worden afgenomen op twee verschillende dagdelen, zodat de leerlingen geconcentreerd aan beide taken kunnen werken. Elk taak bestaat uit 20 opgaven behorende bij een aantal teksten. In totaal bevat elke toets dus 40 opgaven.

Vorm

De toetsen voor groep 4 bevatten een aantal teksten met vragen. Zowel de teksten als de vragen staan in het opgavenboekje. De leerlingen lezen een tekst en beantwoorden daarna een of meer vragen over de tekst.

De opgaven in de toetsen Begrijpend lezen zijn meerkeuzeopgaven. Hierdoor wordt het nakijken en het bepalen van de toetsscore zo eenvoudig en objectief mogelijk gehouden. Elke opgave bevat vier antwoordalternatieven.

Keuze van een passende toets: toetsen op maat

De vaardigheid in begrijpend lezen van leerlingen in een groep loopt vaak sterk uiteen. Als gevolg daarvan zal eenzelfde toets voor begrijpend lezen voor een deel van de leerlingen goed op niveau zijn, maar voor sommige andere leerlingen erg moeilijk of erg gemakkelijk. Met name voor een aantal leerlingen van niveau IV en voor de leerlingen van niveau V (of de leerlingen van niveau D en E) zijn de toetsen van het eigenlijke afnamemoment (bijvoorbeeld de M4-toets voor leerlingen medio groep 4) aan de moeilijke kant. Voor een aantal leerlingen van niveau I (of niveau A) zijn de toetsen echter aan de gemakkelijke kant.

Voor leerlingen die zich minder snel of juist sneller ontwikkelen dan de gemiddelde leerling, is het belangrijk om het niveau van de toets af te stemmen op het niveau van de leerling in plaats van op het aantal jaren onderwijs dat de leerling gevolgd heeft. Dit noemen we toetsen op maat. Zo wordt op de meest betrouwbare manier de vaardigheid van de leerling gemeten. En uiteraard is het maken van een toets op maat prettiger voor de leerlingen. Voor het toetsen op maat wordt gebruikgemaakt van de onderliggende vaardigheidsschaal. Deze schaal maakt het mogelijk om de resultaten van leerlingen die verschillende toetsen voor een bepaald leergebied maken toch met elkaar te vergelijken. Ook kan zo de ontwikkeling van individuele leerlingen in de tijd worden gevolgd. De onderliggende meettechniek voorziet er namelijk in dat iedere ruwe score - op welke toets van Begrijpend lezen deze score ook behaald is - kan worden omgezet in een score op één en dezelfde vaardigheidsschaal. Leerlingen kunnen daardoor bijvoorbeeld een toets maken die hoort bij een vorig afnamemoment (een M4-leerling maakt een toets E3) of een volgend afnamemoment (een M4-leerling maakt de toets E4).

Voor leerlingen met een vertraagde ontwikkeling zijn in de toetsmap voor groep 4 de extra toetsen E3M4 en M4E4 opgenomen.

Afname

De toetsen worden in principe klassikaal afgenomen door de leerkracht of IB'er. De afname start met een klassikale instructie door de leerkracht of IB'er. De toetsmap bevat hiervoor afnamekaarten met afname-instructies aan de hand van een aantal oefenopgaven. Deze oefenopgaven zijn voor de extra toets en de opvolgende reguliere toets hetzelfde. Dat betekent dat de leerkracht/IB'er de klassikale instructie samen kan doen met alle leerlingen die de toets E3M4 en M4 maken. Ditzelfde geldt ook voor de toetsen M4E4 en E4.

Na de instructie maken de leerlingen zelfstandig de toets. Ze lezen de teksten en beantwoorden de vragen door in hun opgavenboekje de letter van hun gekozen antwoord te omcirkelen. De afname is niet aan tijd gebonden. De leerlingen mogen dus de tijd nemen die ze nodig hebben.

De leerkracht/IB'er kan ervoor kiezen om de toets individueel af te nemen bij leerlingen met concentratieproblemen, leerlingen die langzamer dan gemiddeld werken of bij leerlingen die afwezig waren bij de klassikale afname. Belangrijk is dat de leerkracht of IB'er zich ook bij een individuele afname aan de afname-instructies houdt.

In de toetsmappen is een handleiding opgenomen die zich richt op de organisatorische kant van de afname en op de verwerking en interpretatie van de toetsresultaten. In de handleiding is extra aandacht besteed aan het afnemen van de toetsen conform de afname-instructies. Dit is gerealiseerd door aan te geven welke aanpassingen de leerkracht eventueel zelf kan doen en welke invloed dat heeft op de vergelijkbaarheid van de scores.

Scoring

Voor het handmatig nakijken van de toetsen kan gebruikgemaakt worden van een nakijkkaart met goede antwoorden, die in de toetsmap is opgenomen. Indien gewenst kan de leerkracht/IB'er in het Computerprogramma LOVS de opgaven die fout beantwoord zijn aanklikken. Op basis van het aantal goede antwoorden, de toetsscore, wordt een inschatting gemaakt van de vaardigheid van de leerlingen. De leerkracht/IB'er kan ook het aantal goede antwoorden invoeren in het Computerprogramma LOVS. De toetsscore wordt zo automatisch omgezet naar de bijbehorende vaardigheidsscore met een score-interval ofwel betrouwbaarheidsinterval. Een andere optie is om met behulp van de omzettingstabellen in de toetsmap of op Cito Portal de vaardigheidsscore bij de behaalde toetsscore op te zoeken.

Verwerking resultaten en interpretatie

Na de toetsafname en de scoring van de leerlingantwoorden kunnen de toetsresultaten door de leerkracht of IB'er verwerkt worden op speciaal ontwikkelde rapportageformulieren, onder andere leerlingrapporten en groepsoverzichten. Deze rapportages zijn beschikbaar via Cito Portal.

De toetsen kunnen ook met behulp van de computer verwerkt worden. Op schoolniveau kan een IB'er en/of directeur met de computer een dwarsdoorsnede en trendanalyses opvragen. Met behulp van deze overzichten kan de kwaliteit van het gegeven onderwijs op groeps- en schoolniveau geanalyseerd worden. In de handleiding in de toetsmap worden in hoofdstuk 4 de interpretatie- en analysemogelijkheden op leerling- en groepsniveau behandeld. In hoofdstuk 5 van de handleiding komt de interpretatie op schoolniveau aan bod. De handleiding gaat in op de inhoudelijke interpretatie van de rapportages. In de handleiding bij het Computerprogramma LOVS staan de aanwijzingen over de wijze waarop de rapportages zijn op te vragen en welke keuzemogelijkheden de school hierbij heeft.

In de toetsmaterialen zijn twee niveau-indelingen opgenomen, waarmee de leerkracht de scores van een leerling kan vergelijken met die van een grote groep leerlingen (zie ook hoofdstuk 2.3). De leerkracht kan een keuze maken uit een indeling in de niveaus:

- I tot en met V;
- A tot en met E.

Daarnaast heeft de leerkracht de keuze om functioneringsniveaus op te vragen. De functioneringsniveaus geven aan met welke gemiddelde leerling de vaardigheidsscore van de getoetste leerling vergelijkbaar is.

Een functioneringsniveau E4 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling overeenkomt met de score van de gemiddelde leerling eind groep 4. De indeling in functioneringsniveaus is oorspronkelijk ontwikkeld voor het speciaal (basis)onderwijs, om meer inzicht te krijgen in het niveau van de leerlingen met forse leerachterstanden. Mede dankzij de komst van het 'passend onderwijs' ontstond ook bij regulier onderwijs de wens om functioneringsniveaus te gebruiken en zijn de functioneringsniveaus opgenomen in de rapportages.

3.2 Inhoudsverantwoording

In het ontwikkelproces van de toetsen zijn een aantal fasen te onderscheiden:

- uitwerking van de domeinbeschrijving;
- tekstselectie en itemconstructie;
- proeftoetsing en kalibratie-analyses;
- normeringsonderzoek;
- samenstelling van de definitieve toetsen.

We zullen deze fasen hieronder nader toelichten.

Deze informatie vormt een aanvulling op de inhoudsverantwoording die opgenomen is in het toetspakket Begrijpend lezen 3.0 voor groep 4 (hoofdstuk 6). Daarin staan ook specifieke voorbeelden van de verschillende soorten teksten en opgaven die in de toets voorkomen.

3.2.1 Uitwerking domeinbeschrijving in vaardigheden, tekstsoorten en opgavenvormen

De toetsmatrijs voor de toetsen Begrijpend lezen 3.0 is samengesteld op basis van het referentiekader Nederlandse taal (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a), de kerndoelen Nederlandse taal (Ministerie van OCW, 2006), de tussendoelen gevorderde geletterdheid voor de middenbouw (Aarnoutse & Verhoeven, 2003), de Leerstoflijnen lezen (Oosterloo & Paus, 2010) en recente wetenschappelijke publicaties over begrijpend lezen. Op basis van de domeinbeschrijving voor Begrijpend lezen (zie paragraaf 2.4: Theoretische inkadering) zijn de teksten, opgavenvormen en vaardigheden geselecteerd die relevant zijn voor groep 4.

Voorafgaand aan de ontwikkeling van de toetsen Begrijpend lezen zijn de meest gebruikte methoden voor begrijpend lezen bestudeerd, om de tekstsoorten en vraagvormen waar leerlingen in de verschillende jaargroepen in de klas mee werken, in beeld te brengen. Bij de constructie voor de toetsen is, waar mogelijk, rekening gehouden met de 'gemene delers' uit het aanbod, zoals de gemiddelde lengte van teksten, het technisch leesniveau van teksten en de variatie in tekstsoorten, genres, onderwerpen en bronnen. Omdat de leerlijnen van de methodes op hoofdlijnen aan elkaar gelijk zijn, is het onwaarschijnlijk dat het toetsresultaat afhankelijk is van de gevolgde methode.

Indeling in tekstsoorten en tekstgenres

De opgaven in de toetsen Begrijpend lezen hebben in alle gevallen betrekking op teksten of delen van teksten. Deze teksten kunnen worden onderverdeeld in twee hoofdcategorieën: zakelijke teksten enerzijds en literaire en fictionele teksten anderzijds (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a). Voor de toetsen Begrijpend lezen zijn binnen deze hoofdcategorieën vijf tekstsoorten onderscheiden:

- *Informatieve teksten*
De schrijver geeft feitelijke informatie over de werkelijkheid;
- *Instructieve teksten*
De schrijver wil het gedrag of het handelen van de lezer richten en sturen;
- *Betogende teksten*
De schrijver wil het denken of het handelen van de lezer beïnvloeden;
- *Verhalende teksten*
De schrijver beschrijft een verbeelde werkelijkheid in verhalen;
- *Poëzieteksten*
De schrijver beschrijft een verbeelde werkelijkheid in poëzie.

De informatieve, instructieve en betogende teksten nemen we verderop samen onder de noemer 'formele teksten' ofwel zakelijke teksten. De verhalende teksten en poëzieteksten nemen we samen onder de noemer 'informele teksten'.

Bij het onderscheid tussen tekstsoorten, hierboven beschreven, gaat het om functies of doelen van teksten: met een betogende tekst wil de schrijver de lezer bijvoorbeeld ergens van overtuigen, met een verhalende tekst wil de schrijver lezers vermaken of boeien, enzovoort. Teksten kunnen ook op grond van een ander criterium onderscheiden worden. Een tekst kan een brief zijn, maar ook een verslag, een recept, enzovoort. Dit onderscheid wordt omschreven als het tekstgenre. Tekstgenres waar basisschoolleerlingen mee te maken kunnen krijgen zijn onder andere: verhaal, gedicht, artikel, uitnodiging, verslag, brief, instructie. Soms hebben tekstgenre en tekstsoort een een-op-een-relatie: een nieuwsbericht is altijd informatief, een recept is altijd instructief. Soms ligt de relatie echter complexer: een artikel kan informatief zijn, maar er kan ook iets in beargumenteerd worden en er kunnen alle mogelijke aanwijzingen in gegeven worden. Daarnaast kan een tekst onder meerdere genres vallen: een tekst kan bijvoorbeeld de vorm van een brief hebben, maar kan ook een verslag zijn.

De verdeling van tekstsoorten en tekstgenres in de toetsen Begrijpend lezen is gebaseerd op het dagelijks aanbod aan leesteksten dat de leerlingen op school en thuis tegenkomen. Een groot deel van de teksten in de toetsen zou daarom moeten bestaan uit (delen van) verhalen uit leesboeken en een ander groot deel uit eenvoudige informatieve teksten. Naarmate het toetsniveau hoger wordt, komt het accent wat minder op informele teksten (met name verhalende teksten) te liggen en wat meer op formele teksten (informatieve, instructieve en betogende teksten). Dit komt overeen met het aanbod aan type teksten waarmee leerlingen gedurende het basisonderwijs geconfronteerd worden.

Voor alle toetsen van groep 4 was het streven dat de helft van de teksten informeel zou zijn en de helft formeel (zakelijk). De informele teksten zouden uit verhalende teksten bestaan en hooguit één poëzietekst. De formele teksten zouden voor het grootste deel uit informatieve teksten moeten bestaan. Daarnaast zouden instructieve en betogende teksten opgenomen kunnen worden. Voor de tekstgenres was het streven dat er zoveel mogelijk diversiteit is.

Indeling in opgavenvormen

De opgaven in de toetsen Begrijpend lezen zijn onder te brengen in een aantal categorieën, gelet op de vorm van de opgaven. Om te beginnen zijn alle opgaven in deze toetsen meerkeuzeopgaven. Hierdoor wordt het nakijken van en het bepalen van de scores op de toetsen zo eenvoudig en zo objectief mogelijk gehouden. Het nakijken van open opgaven vereist scoringsvoorschriften, waarmee het aanzienlijk lastiger werken is dan met de objectief toepasbare toetssleutels in de vorm van lijsten met goede antwoorden, zoals die staan in de handleidingen bij de toetsen Begrijpend lezen.

Traditioneel zijn opgaven Begrijpend lezen van de vorm 'vragen over teksten' (tekstopgaven). Een opgave uit deze categorie kan omschreven worden als: een vraag die gesteld wordt naar aanleiding van een tekst of een deel van een tekst (een of meer zinnen, een of meer alinea's, enzovoort). Het is de opgavenvorm die ook in de methoden voor begrijpend lezen het vaakst gehanteerd wordt.

Daarnaast worden in Begrijpend lezen 3.0 de opgavenvormen 'openplaatsopgaven' en 'voorspelopgaven' onderscheiden. Bij openplaatsopgaven worden teksten aangeboden waaruit een zin, zinsdeel of woord is weggelaten. De leerlingen kiezen het alternatief dat het best in de tekst past, door zowel het stuk vóór de invulplaats, als het stuk na de invulplaats te lezen. Met de openplaatsopgaven wordt het leesbegrip op een andere manier bevraagd dan met de tekstopgaven. Alleen als de leerling het stukje rond de open plaats goed heeft begrepen, kan hij het juiste antwoord kiezen. Voorspelopgaven zijn opgaven waarbij alleen de titel, een (begin)gedeelte van de tekst en (soms) een afbeelding worden gegeven. De rest van de tekst ontbreekt. Leerlingen moeten op basis van de gegeven informatie een opgave beantwoorden waarbij ze een voorspelling moeten doen over de inhoud van de tekst, de bron van de tekst of de tekstsoort. Deze opgavenvorm sluit aan bij hedendaagse methoden voor begrijpend lezen waarbij leerlingen leren om voorafgaand aan het lezen van een tekst te voorspellen waar de tekst over zal gaan.

Bij de verdeling van de opgavenvormen over de toetsen Begrijpend lezen was het uitgangspunt dat het zwaartepunt van de toets uit vragen bij teksten moet bestaan, aangezien dit ook in de lesmethoden de meest bekende en meest gebruikte vorm van leesopgaven is. Openplaatsopgaven en voorspelopgaven zouden een kleiner deel van de toets moeten beslaan. Het aandeel openplaatsopgaven zou moeten afnemen naarmate het toetsniveau hoger is, omdat deze opgavenvorm in de hogere leerjaren al snel te makkelijk is. Het aandeel tekstopgaven kan dan groter worden.

Voor de toetsen E3M4 en M4 was het streven dat er 40% openplaatsopgaven opgenomen werden, 50% tekstvragen en 10% voorspelopgaven. Voor de toetsen M4E4 en E4 was het streven dat er 30% openplaatsopgaven, 60% tekstopgaven en 10% voorspelopgaven opgenomen werden.

Indeling in vaardigheden

Bij het construeren van de opgaven onderscheidde we twee (deel)vaardigheden van begrijpend lezen: het begrijpen van geschreven teksten en het interpreteren van geschreven teksten. De vaardigheid begrijpen heeft betrekking op de verwerking van informatie die expliciet in de tekst vermeld staat. Bij interpreteren gaat het erom dat de lezer de informatie uit de tekst verbindt aan zijn eigen kennis, waaronder zijn kennis van de wereld en kennis van taal en tekstkenmerken. De twee vaardigheden staan niet op zichzelf, maar liggen in elkaars verlengde en worden in wisselwerking met elkaar toegepast tijdens het lezen van een tekst.

Binnen de vaardigheden begrijpen en interpreteren worden opgaven verder onderverdeeld. Daarbij is het gedeelte van de tekst waarop de opgave betrekking heeft het uitgangspunt. Een opgave kan bijvoorbeeld gebaseerd zijn op een of meerdere zinnen, een alinea, of op de tekst als geheel.

Opgaven bij Begrijpen van geschreven taal

- 1 Opgaven die vragen naar de betekenis van een woord, woordgroep of zin(nen) die expliciet in de tekst vermeld wordt. (categorie B1).
- 2 Opgaven die vragen naar specifieke inhoudselementen die expliciet in de tekst aan de orde gesteld worden. Dit zijn bijvoorbeeld feiten en meningen, voorwerpen, aantallen, een plaats van handeling of tijdsperiodes, (hoofd)personen. (categorie B2).
- 3 Opgaven die vragen naar eenvoudige expliciete verbanden op lokaal niveau. Verbanden kunnen worden gelegd op basis van inhoudelijke en/of structurele elementen zoals signaalwoorden. Voorbeelden van verbanden zijn verwijzingen, vergelijkingen, tegenstellingen, generalisaties en voorbeelden, oorzaak en gevolg, vraag en antwoord, reden en verklaring, middel en doel, deel-/geheelrelaties, conclusie en argumenten. (categorie B3).
- 4 Opgaven die vragen naar complexe expliciete verbanden over grotere tekstdelen heen. Verbanden kunnen worden gelegd op basis van inhoudelijke en structurele elementen. Voorbeelden van verbanden zijn vergelijkingen, tegenstellingen, volgordes, generalisaties en voorbeelden, oorzaak en gevolg, vraag en antwoord, reden en verklaring, middel en doel, deel-/geheelrelaties, conclusie en argumenten. (categorie B4).

Opgaven bij Interpreteren van geschreven taal

De indeling van de opgaven bij Interpreteren van geschreven taal is als volgt:

- 1 Opgaven die vragen naar het afleiden van de betekenis van een woord, woordgroep of zin(nen). (categorie I1)
- 2 Opgaven die vragen naar het afleiden van informatie uit de tekst op lokaal niveau. De leerling moet zijn voorkennis inzetten naast de informatie die de tekst geeft. Voorbeelden hiervan zijn opgaven over onderwerp, thema, hoofdlijnen, hoofdgedachte, hoofdpersoon, setting en doel en doelgroep van de tekst, het vertelperspectief en de tekststructuur. (categorie I2)

- 3 Opgaven die vragen naar het afleiden van informatie op het globale niveau van de tekst. Verbanden kunnen worden gelegd op basis van inhoudelijke en/of structurele elementen. Voorbeelden hiervan zijn opgaven over onderwerp, thema, hoofdlijnen, hoofdgedachte, hoofdpersoon, setting en doel en doelgroep van de tekst, het vertelperspectief en de tekststructuur. (categorie I3)
- 4 Opgaven die vragen naar het taalgebruik en schrijfstijl van de schrijver en waar verbanden moeten worden gelegd tussen tekstuele informatie en kennis van het taalsysteem. Voorbeelden hiervan zijn vragen naar tekstsoort, naar het gebruik van aanhalingstekens of vragen naar specifieke woordkeuze waarbij register, sociale en culturele conventies een rol spelen. (categorie I4)

De verschillende categorieën kunnen in alle jaargroepen voorkomen, maar complexiteit van de vorm en inhoud van de opgaven binnen de categorieën neemt toe voor de hogere groepen.

De categorieën B2 tot en met B4 en I2 tot en met I4 zijn onderverdeeld in verschillende subcategorieën. De categorieën B1 en I1 bevatten slechts één subcategorie. Omdat de categorieën B1 en I1 zo smal zijn, is de verwachting dat die incidenteel voorkomen. Het is niet per definitie noodzakelijk dat elke toets opgaven van deze categorieën bevat.

De indeling in categorieën was vooral van belang bij de constructie van de opgaven, zodat de constructeurs er alert op waren om de vaardigheden Begrijpen en Interpreteren in de volle breedte te bevragen. Voor de selectie van de uiteindelijke opgaven spelen ze een minder belangrijke rol. Daar spelen ook de tekstsoorten en opgavenvormen een grote rol. Een belangrijke voorwaarde voor de vaardigheden Begrijpen en Interpreteren is dat de beide vaardigheden - inclusief het merendeel van de categorieën (of subvaardigheden) waarin deze vaardigheden zijn opgedeeld -, gedekt zijn in een toets.

Aspecten van begrijpend lezen

Hierboven werden de tekstsoorten, opgavenvormen en (deel)vaardigheden, die in de toetsen Begrijpend lezen onderscheiden worden, omschreven. In paragraaf 3.1 van deze verantwoording is de schaal Begrijpend lezen geïntroduceerd. Op deze meetschaal kan de leesvaardigheid van leerlingen afgebeeld worden, maar ook kunnen de opgaven van de toetsen Begrijpend lezen er naar moeilijkheid op gerangschikt worden.

Alle opgaven hebben een plaats op deze ene schaal. De opgaven van de toetsen Begrijpend lezen vormen, zoals dat heet, een unidimensionele meetschaal: de opgaven meten een en dezelfde vaardigheid. Een vaardigheid waaraan in het primair onderwijs de naam Begrijpend lezen is gegeven. Het is niet zo dat de opgaven van één (sub)type systematisch op een bepaald deel van de schaal liggen en de opgaven van een ander (sub)type op een ander deel van de schaal. Uit de schaal kan dus ook niet afgeleid worden dat bijvoorbeeld het tweede (sub)type opgaven moeilijker is dan het eerste (sub)type. Het is dan ook niet zinvol om scores per opgaventype te rapporteren als een soort deeltoets, en om deze scores van een leerling te presenteren als een niveau-indicatie van een onderliggende deelvaardigheid van het begrijpend lezen. De opgaven zijn desondanks ingedeeld in verschillende deelvaardigheden en opgavenvormen, omdat een dergelijke indeling in de ontwikkelingsfase van de toetsen – met name tijdens het construeren van de opgaven – ervoor zorgt dat de vaardigheid Begrijpend lezen in al zijn facetten en van alle kanten belicht wordt.

De toets Begrijpend lezen brengt tekstbegrip in het algemeen in beeld. Om de vragen in de toetsen te kunnen beantwoorden moeten de leerlingen strategieën toepassen voorafgaand aan het lezen, tijdens en na het lezen. Verschillende strategieën zijn hierbij mogelijk om tot tekstbegrip te komen. Met de toetsen Begrijpend lezen meten we niet welke strategie de leerling heeft toegepast, maar meten we of de leerling de tekst heeft begrepen. Het toepassen van strategieën is namelijk een hulpmiddel en geen doel op zich bij leesbegrip.

3.2.2 Itemconstructie, onderzoeken en selectie van opgaven voor de toetsen Begrijpend lezen

Tekstselectie en itemconstructie

Alle opgaven die in de toetsen Begrijpend lezen zijn opgenomen werden speciaal voor deze toetsen geconstrueerd door een constructiegroep, bestaande uit leerkrachten basisonderwijs, schoolbegeleiders en pabodocenten. De constructeurs kregen een opdracht, die was opgesteld door toetsdeskundigen van Cito.

In deze opdracht stond omschreven wat voor type teksten gezocht moesten worden en wat voor opgaventypen erbij gemaakt moesten worden. De constructeurs hadden van de toetsdeskundigen uitgebreide richtlijnen voor constructie ontvangen, zowel op papier als mondeling. De richtlijnen bevatten onder andere aanwijzingen voor de selectie van teksten, de constructie van de verschillende opgavenvormen, de formulering van de vraagstelling en de alternatieven en het gebruik van illustraties. De geselecteerde teksten en de geconstrueerde items zijn in constructiegroepvergaderingen onder leiding van een toetsdeskundige besproken en vervolgens bijgesteld. Een illustrator kreeg van toetsdeskundigen van Cito richtlijnen voor illustraties.

Proeftoetsing en kalibratie-analyses

De opgaven zijn eerst in proefafnames voorgelegd aan leerlingen in de jaargroepen waarvoor ze bedoeld waren. Het doel van dergelijke proefafnames is het verkrijgen van informatie over de moeilijkheid van elke opgave. Tevens kunnen eventuele slecht functionerende opgaven (bijvoorbeeld opgaven die vaker door goede lezers dan door minder goede lezers fout gemaakt worden) geïdentificeerd en verwijderd worden. Daarnaast wilden we in de proeftoetsing nagaan hoe de leerkrachten de teksten en opgaven ervoeren. De leerkrachten die aan de proeftoetsing deelnamen, hebben we gevraagd een evaluatieformulier in te vullen. Hierop konden ze onder andere inhoudelijke opmerkingen maken en aangeven welke teksten of opgaven ze minder geschikt of ongeschikt vonden.

In het proefonderzoek M4 zijn 415 items voorgelegd aan 2550 leerlingen van groep 4 verdeeld over 26 boekjes. Elk boekje bestond uit 48 tot 51 opgaven verdeeld over 2 taken. Een van de twee taken betrof de starttaak uit LVS 2.0. Deze taak werd door alle 2550 leerlingen gemaakt. De overige opgaven kwamen in een of twee boekjes voor en werden gemiddeld door 168 leerlingen gemaakt.

In het proefonderzoek E4 zijn 396 items voorgelegd aan 2676 leerlingen van groep 4 verdeeld over 20 boekjes. Elk boekje bestond uit 47 tot 51 opgaven verdeeld over 2 taken. Een van de twee taken betrof de starttaak uit LVS 2.0. Deze taak werd door alle 2676 leerlingen gemaakt. De overige opgaven kwamen in een of twee boekjes voor en werden gemiddeld door 179 leerlingen gemaakt.

Na de afnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programmapakket One Parameter Logistic Model (OPLM; Verhelst, 1993; Verhelst en Glas, 1995). Zie voor een algemene technische beschrijving van dit model paragraaf 2.4.2.

Bij de analyses is de kwaliteit van de afzonderlijke items en de totale verzameling voor een afnamemoment in kaart gebracht. Itemparameters en discriminatieparameters zijn geschat. Bij de analyses van de antwoorden van de leerlingen op de opgaven is nagegaan of de verschillende onderdelen een beroep doen op hetzelfde complex aan vaardigheden. Dat bleek (voor de meeste opgaven) het geval te zijn; opgaven die niet voldeden vielen af.

Na de uitwerking van de opgaven door toetsdeskundigen van Cito zijn de opgaven nogmaals gescreend door praktijkdeskundigen uit het SBO. Hierbij is erop gelet dat de opgaven geschikt zijn voor een zo groot mogelijke groep leerlingen, ook leerlingen met extra onderwijsbehoeften. Deze screening heeft er niet toe geleid dat teksten of opgaven verwijderd werden. Men vond ze ook geschikt voor het SBO. Wel gaven de praktijkdeskundigen aan bepaalde opgaven moeilijk te vinden, maar dat was niet specifiek voor SBO-leerlingen. Dat bleken inderdaad moeilijke opgaven te zijn voor het merendeel van de leerlingen; deze moeten erin blijven om de goede leerlingen van de zwakke leerlingen te kunnen onderscheiden.

Normeringsonderzoek

Op basis van de psychometrische analyses en de evaluaties van de proeftoetsing hebben we de teksten en opgaven geselecteerd voor het normeringsonderzoek.

De psychometrische criteria betroffen met name de moeilijkheidsgraad en discriminatieparameter. Waar mogelijk hebben we bij de selectie rekening gehouden met opmerkingen die de leerkrachten gemaakt hebben bij bepaalde teksten en opgaven.

Voor een evenwichtige samenstelling van de taken hebben we vervolgens gekeken naar de gewenste indelingen voor tekstsoorten, opgavenvormen en vaardigheden zoals die in paragraaf 3.2.1 staan. Deze gewenste indelingen golden als richtlijnen bij de selectie.

Voor de indeling van teksten en items over de taken hebben we vervolgens gelet op de volgende punten:

- Per taak afwisselende onderwerpen van de teksten.
- Per taak afwisselende tekstgenres.
- Per taak afwisselende lengte van de teksten.
- Zelfde volgorde van opgavenvormen: elke taak begint met openplaatsopgaven, daarna komen de voorspelopgaven en tenslotte de tekstopgaven.
- Elke taak begint en eindigt met relatief makkelijke items.
- De taken hebben een gemiddelde p-waarde van rond de .70.

De taken in het normeringsonderzoek leken dus al zoveel mogelijk op de definitieve taken. Wel bevatte elke taak enkele opgaven méér dan de definitieve taken, met het oog op eventuele uitval van items wegens slecht functioneren.

Samenstelling definitieve toetsen

Na het normeringsonderzoek zijn nog enkele opgaven verwijderd om tot de gewenste aantallen te komen. Bij de samenstelling van de toetsen is rekening gehouden met een mogelijk volgorde-effect: de volgorde van de opgaven in de definitieve toetsen wijkt minimaal af van die van het normeringsonderzoek. Onderzoek naar de beperkte wijzigingen in volgorde liet zien dat de wijziging in de volgorde geen effect had.

De opgaven met een acceptabele moeilijkheid (in klassieke termen: p-waarde tussen .40 en .90) die door de betere lezers significant vaker goed werden gemaakt dan door de minder goede lezers (r_{ir} vanaf .20) kwamen in principe in aanmerking voor opname in de definitieve toetsen Begrijpend lezen.

Bij het selecteren van de opgaven speelden echter ook inhoudelijke criteria. Zoals hierboven al vermeld, moesten de uiteindelijke toetsen een evenwichtige verzameling teksten en opgaven bevatten. Hierbij werd gelet op tekstsoort, tekstgenre, opgavenvorm en de inhoud (het onderwerp) van de teksten. Daarnaast is ervoor gezorgd dat het aantal opgaven in verhouding staat tot de tekst waarbij de opgaven horen.

In de meeste gevallen kon aan zowel de psychometrische (p-waarde, r_{ir}) als de inhoudelijke criteria voldaan worden. In sommige gevallen zijn er op basis van de inhoudelijke criteria opgaven opgenomen die (net) niet de gewenste psychometrische waarden hadden.

Sommige opgaven die afvielen, vertoonden in het normeringsonderzoek een te hoge of te lage moeilijkheid of een te laag discriminerend vermogen. Soms vielen opgaven af die psychometrisch gezien goed functioneerden, maar die op inhoudelijke gronden afgewezen werden (hiervoor werden dezelfde criteria gehanteerd als bij de proefafname). Daarentegen werden soms opgaven gehandhaafd die eigenlijk wat te moeilijk of te makkelijk waren, maar die wel zorgen voor een inhoudelijk beter samengestelde verzameling opgaven en teksten. Bij elke verzameling opgaven bij een tekst vond dus een afweging plaats op zowel psychometrische als inhoudelijke gronden.

De toetsen bevatten opgaven van uiteenlopende moeilijkheidsgraad. De toetsen zijn hierdoor geschikt om verschillen tussen leerlingen in beeld te brengen. Een goede illustratie hiervan en van de samenstelling van de toetsen zijn de figuren in bijlage 4: p50 en p80-kanspunten van de opgaven in de toetsen voor groep 4 in relatie tot de gemiddelde vaardigheidsscore voor de afnamemomenten. In deze figuren is zichtbaar dat de toetsen opgaven bevatten van uiteenlopende moeilijkheidsgraad. In de figuren is de verdeling van de opgaven over de taken van de toetsen visueel weergegeven. De balkjes in de figuren geven het p50- (onderkant van het balkje) en p80-kanspunt (bovenkant van het balkje) van elke opgave aan. Het p50-punt geeft de vaardigheidsscore aan waarbij er sprake is van een kans van 50% om een opgave goed te beantwoorden.

Bij de toets M4 ligt het merendeel van de balkjes op en rond de gemiddelde vaardigheidsscore behorende bij medio groep 4. Er is een goede spreiding in moeilijkheidsgraad van de opgaven: er zijn makkelijke opgaven (die liggen onder de lijn van M4), opgaven van gemiddelde moeilijkheid (doorkruisen de lijn van M4) en (erg) moeilijke opgaven (liggen boven de lijn van M4). Bij deze toets kunnen ook de betere leerlingen laten zien wat ze in huis hebben. Bij de gemakkelijke variant van de toets M4, de toets E3M4, liggen vrijwel alle balkjes onder de lijn van M4 (dus onder de gemiddelde vaardigheidsscore). Uit de figuren van E3M4 is duidelijk op te maken dat deze toets over de gehele linie erg makkelijk is en dus geschikt voor de zwakkere lezers. Bij de toetsen M4E4 en E4 liggen de opgaven veelal hoger op de vaardigheidsschaal,

rond de gemiddelde vaardigheidsscore eind groep 4, waarbij de toets E4 weer moeilijker is dan de toets M4E4. Ook is te zien dat de groei in vaardigheid tussen de momenten M4 en E4 maar klein is: de lijntjes van de gemiddelde vaardigheidsscores M4 resp. E4 liggen vrij dicht op elkaar.

In alle figuren is zichtbaar dat de toetsen relatief veel gemakkelijke opgaven bevatten, oftewel opgaven behorend bij lagere vaardigheidsscores dan de gemiddelde vaardigheidsscore op het betreffende toetsmoment. Deze keuze is gemaakt om te zorgen dat het merendeel van de leerlingen een succeservaring heeft bij het maken van de toets. Er is gezocht naar een optimale balans tussen nauwkeurig uiteenlopende vaardigheidsniveaus in beeld brengen en zorgen voor een prettige toetservaring voor leerling en leerkrachten.

3.2.3 Gerealiseerde verdeling van toetsitems

Verdeling van tekstsoorten

Zoals in paragraaf 3.2.1 staat, was het streven dat de uiteindelijke toetsen voor groep 4 voor de helft uit informele en voor de helft uit formele (zakelijke) teksten zou bestaan. De informele teksten zouden uit verhalende teksten bestaan en hooguit één poëzietekst. De formele teksten zouden voor het grootste deel uit informatieve teksten moeten bestaan. Daarnaast zouden instructieve en betogende teksten opgenomen kunnen worden.

In tabel 3.1 is te zien dat bovenstaande helemaal geldt voor de toets M4. De toets E3M4 heeft nog iets meer informele teksten (in de lijn van de toets E3), de toetsen M4E4 en E4 hebben iets meer formele teksten. Het is onmogelijk om precies de gewenste verdelingen te krijgen in de uiteindelijke toetsen, omdat we moeten selecteren uit de overgebleven items na de proeftoets en na het normeringsonderzoek. Naast de tekstsoorten, moeten we uiteraard ook rekening houden met de psychometrische kwaliteit, de verdelingen van opgavenvormen en de opmerkingen van leerkrachten. Gezien deze beperkingen is de verdeling van tekstsoorten goed gelukt.

Tabel 3.1 *Tekstsoorten in Begrijpend lezen groep 4: percentage teksten (aantal teksten)*

Tekstsoort	formeel			informeel		Totaal aantal teksten
	Informatief	Instructief	Betogend	Verhalend	Poëzie	
Toets						
E3M4	20% (3)	7% (1)	13% (2)	53% (8)	7% (1)	15
M4	29% (4)	14% (2)	7% (1)	43% (6)	7% (1)	14
M4E4	40% (6)	13% (2)	7% (1)	33% (5)	7% (1)	15
E4	25% (4)	19% (3)	13% (2)	38% (6)	6% (1)	16

Verdeling van tekstgenres

Voor de tekstgenres was het streven dat er zoveel mogelijk diversiteit is. In tabel 3.2 is te zien dat het aantal tekstgenres varieert van 5 (in de toets M4) tot 9 (in de toets E4). Hiermee zijn de toetsen behoorlijk gevarieerd geworden. Sommige teksten bleken overigens in te delen bij meerdere genres: zo waren er artikelen waar ook instructies/aanwijzingen in werden gegeven. In zulke gevallen zijn de teksten ingedeeld bij het genre waar ze de meeste kenmerken mee deelden.

Tabel 3.2 Tekstgenres in Begrijpend lezen groep 4: percentage teksten (aantal teksten)

Toets \ Tekstgenre	E3M4	M4	M4E4	E4
Artikel	20% (3)	29% (4)	33% (5)	19% (3)
Uitnodiging	7% (1)			
Instructie	7% (1)	14% (2)	13% (2)	6% (1)
Brief	7% (1)			
Aankondiging		7% (1)		6% (1)
(Nieuws)bericht			7% (1)	6% (1)
Oproep				6% (1)
Recept				13% (2)
Verslag			7% (1)	6% (1)
Verhaal	53% (8)	43% (6)	33% (5)	31% (5)
Gedicht	7% (1)	7% (1)	7% (1)	
Lied				6% (1)
Totaal aantal teksten	15	14	15	16

Verdeling van opgavenvormen

Zoals aangegeven in paragraaf 3.2.1 was het uitgangspunt dat het zwaartepunt van de toetsen Begrijpend lezen uit vragen bij teksten moet bestaan, aangezien dit ook in de lesmethoden de meest bekende en meest gebruikte vorm van leesopgaven is. Openplaatsopgaven en voorspelopgaven zouden een kleiner deel van de toets moeten beslaan.

Voor de toetsen E3M4 en M4 was het streven dat er 40% openplaatsopgaven opgenomen werden, 50% tekstvragen en 10% voorspelopgaven. Voor de toetsen M4E4 en E4 was het streven dat er 30% openplaatsopgaven, 60% tekstopgaven en 10% voorspelopgaven opgenomen werden.

In tabel 3.3 is te zien dat het aandeel openplaatsopgaven in de toetsen minder is dan vooraf gewenst. Dit heeft vooral te maken met psychometrische overwegingen. Veel van de openplaatsopgaven bleken ofwel veel te moeilijk ofwel veel te makkelijk te zijn.

Tabel 3.3 Opgavenvormen in Begrijpend lezen groep 4: percentage opgaven (aantal opgaven)

Toets \ Opgavenvorm	Openplaatsopgave	Vraag over tekst	Voorspelopgave	Totaal aantal opgaven
E3M4	30% (12)	60% (24)	10% (4)	40
M4	32.5% (13)	60% (24)	7.5% (3)	40
M4E4	27.5% (11)	62.5% (25)	10% (4)	40
E4	25% (10)	67.5% (27)	7.5% (3)	40

Verdeling van vaardigheden

Bij de toetsconstructie voor de toetsen Begrijpend lezen was het uitgangspunt dat de verschillende categorieën (zie paragraaf 3.2.1) in alle jaargroepen konden voorkomen en dat de complexiteit van de vorm en inhoud van de opgaven binnen de categorieën zou toenemen voor de hogere groepen. Een belangrijke voorwaarde voor de vaardigheden Begrijpen en Interpreteren is dat de beide vaardigheden – inclusief het merendeel van de categorieën (of subvaardigheden) waarin deze vaardigheden zijn opgedeeld –, gedekt zijn in een toets. In tabel 3.4 is te zien dat deze wens gerealiseerd is.

Tabel 3.4 Vaardigheden in Begrijpend lezen groep 4: percentage opgaven (aantal opgaven)

Vaardigheid	B1	B2	B3	B4	I1	I2	I3	I4	Totaal aantal opgaven
Toets									
E3M4		10% (4)	15% (6)	2.5% (1)	2.5% (1)	40% (16)	27.5% (11)	2.5% (1)	40
M4		2.5% (1)	17.5% (7)	10% (4)		57.5% (23)	12.5% (5)		40
M4E4		5% (2)	15% (6)	5% (2)		47.5% (19)	25% (10)	2.5% (1)	40
E4		5% (2)	12.5% (5)	2.5% (1)		45% (18)	32.5% (13)	2.5% (1)	40
E3M4		10% (4)	15% (6)	2.5% (1)	2.5% (1)	40% (16)	27.5% (11)	2.5% (1)	40

3.3 Statistische beschrijving

In hoofdstuk 4 zullen de kalibratie en normering uitgebreid worden beschreven. Voorafgaand aan deze uitgebreide beschrijving geven we hier een globaal overzicht van de toetsen E3M4, M4, M4E4 en E4. Opgemerkt dient te worden dat de toets E3M4 als gemakkelijke versie voor moment M4 wordt gebruikt en de toets M4E4 als gemakkelijke versie voor moment E4. De toets E3M4 is afgenomen in zowel de normeringsonderzoeken E3 als M4, terwijl de toets M4E4 zowel in normeringsonderzoeken M4 als E4 is afgenomen (zie ook design hoofdstuk 4). In tabel 3.5 worden de beschrijvende gegevens van de toetsen E3M4, M4, M4E4 en E4 gegeven, zowel op de ruwe scoreschaal als op de vaardigheidsschaal. Omdat de tussentoetsen als makkelijke versie van het afnamemoment erboven worden gezien, zijn de gegevens op de vaardigheidsschaal gelijk. De gegevens zijn voor E3M4 en M4 gebaseerd op 3350 leerlingen en voor M4E4 en E4 op 4859 leerlingen.

Tabel 3.5 Beschrijvende gegevens Toetsen E3M4, M4, M4E4 en E4 op ruwe score en vaardigheidsschaal

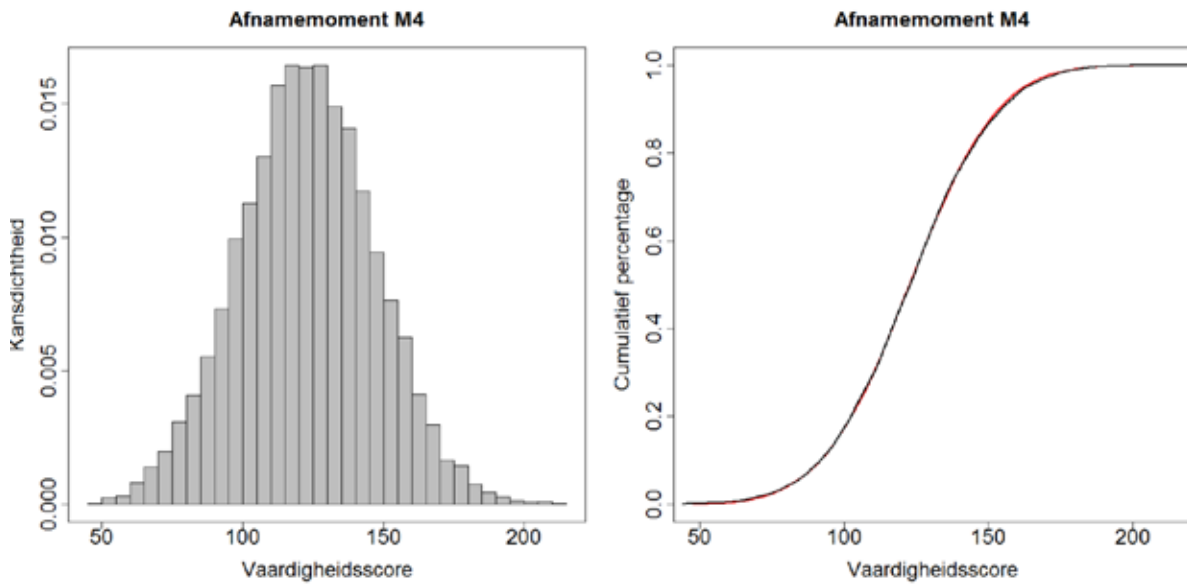
Score	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
E3M4 Ruwe score	31.7	7.13	1.235	-1.252
E3M4 Vaardigheid	133.8	25.48	-.042	.035
M4 Ruwe score	28.2	6.91	.101	-.745
M4 Vaardigheid	133.8	25.48	-.042	.035
M4E4 Ruwe score	29.4	6.79	.405	-.889
M4E4 Vaardigheid	138.8	24.92	-.107	.061
E4 Ruwe score	28.1	7.45	.025	-.744
E4 Vaardigheid	138.8	24.92	-.107	.061

De ruwe scores zijn licht scheef verdeeld. Dit is niet verwonderlijk. De toetsen worden samengesteld rond een verwachte (gewenste) p-waarde van .70 voor de reguliere versies M4 en E4 en een makkelijkere versie (gewenste p-waarde .75-.80 voor de tussentoetsen E3M4 en M4E4.

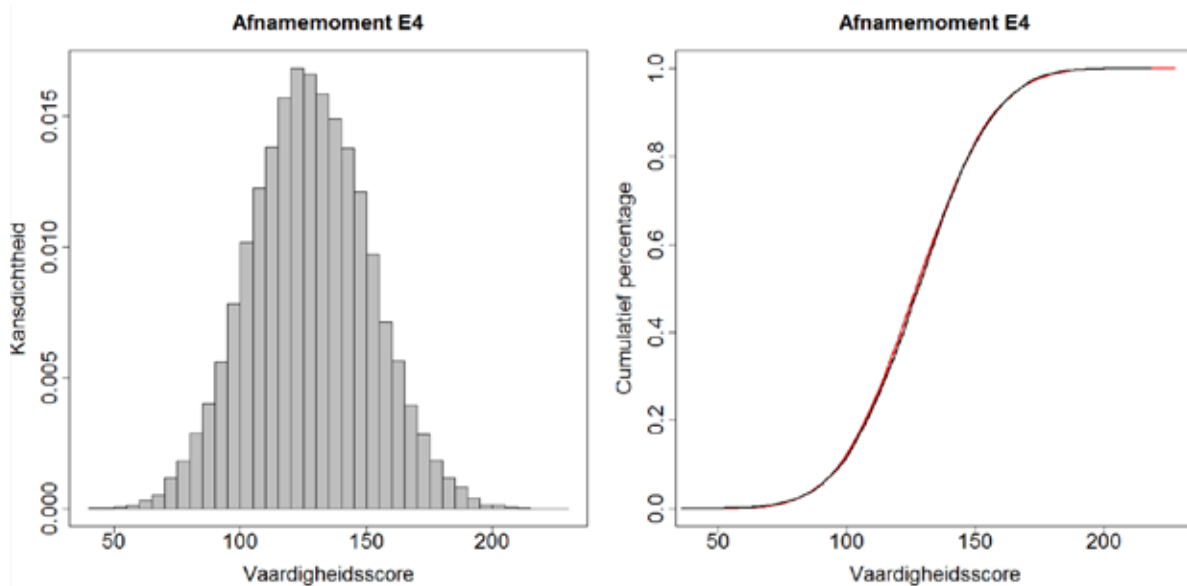
De vaardigheidsverdeling, de scores die gebruikt worden om leerlingen te vergelijken en te volgen, is echter wel normaal verdeeld, zoals te zien is in het linkerpaneel van figuur 3.1 en figuur 3.2. De rechter figuur geeft de cumulatieve verdeling weer (in rood) alsmede de cumulatieve verdeling behorend bij de normaalverdeling (in zwart). Deze verdelingen ontlopen elkaar zeer weinig.

Het verschil tussen de vaardigheidsscores zoals weergegeven in figuur 3.1 en figuur 3.2 en die in tabel 3.5 is te verklaren doordat de scores in figuur 3.1 en figuur 3.2 worden weergegeven op de kalibratieschaal voor groep 4 terwijl de gemiddelden en standaarddeviaties in tabel 3.3 op de (getransformeerde) overkoepelende vaardigheidsschaal worden weergegeven, zie ook hoofdstuk 4.

Figuur 3.1 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling afnamemoment M4



Figuur 3.2 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling E4



4 Kalibratie en normering

4.1 Opzet voor de normeringsonderzoeken van het LVS: het macrodesign

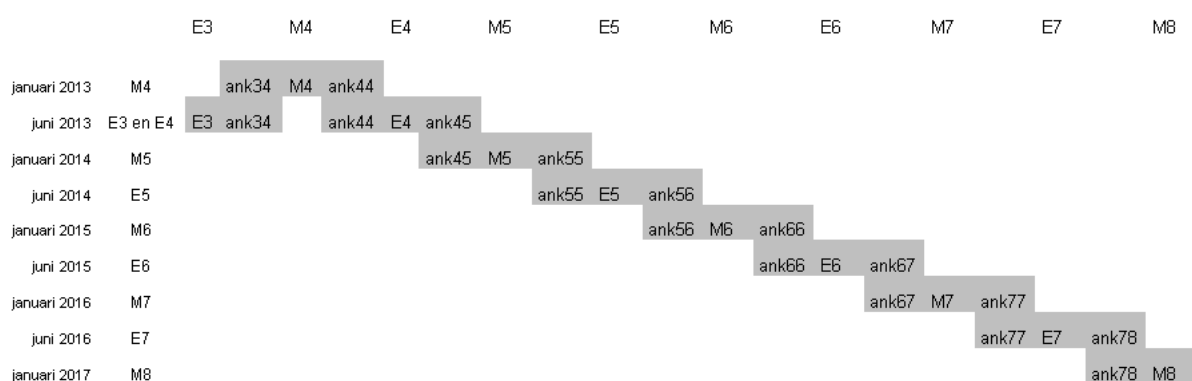
Het opzetten van een leerlingvolgsysteem in het basisonderwijs is een complexe onderneming, en het verzamelen van de gegevens om het systeem te ijken en normeren moet met de nodige zorg gebeuren. Immers, het is niet voldoende om voor elke halfjaargroep (E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8) over normen te beschikken, er moet ook voor gezorgd worden dat de prestaties over de jaren heen met elkaar vergelijkbaar zijn. Hiertoe dienen de prestaties van leerlingen over alle leerjaren heen te worden afgebeeld op een gemeenschappelijke vaardigheidsschaal. Om zo'n gemeenschappelijke schaal te realiseren kunnen we niet volstaan met het ontwikkelen van afzonderlijke toetsen voor de meetmomenten en elke toets afzonderlijk ijken en normeren. Prestaties van bijvoorbeeld de populatie M5 moeten vergelijkbaar zijn met die van andere populaties, bijvoorbeeld E4 en E5, oftewel het dataverzamelingsdesign, dient verbonden te zijn. Hiertoe dient een longitudinale opzet gebruikt te worden.

De verbondenheid van het design

Het idee van een gemeenschappelijke schaal impliceert strikt genomen dat men iemands vaardigheid zou kunnen schatten aan de hand van een willekeurig samengestelde toets. Het spreekt echter vanzelf dat het een zinloze onderneming is een toets die geconstrueerd is voor groep 7 voor te leggen aan leerlingen van groep 3, omdat zo'n toets ongetwijfeld opgaven zal bevatten die een beroep doen op kennis van leerstof die in groep 3 niet is onderwezen. Dit betekent dat we door de algemene kenmerken van het curriculum tamelijk beperkt zijn in het voorleggen van itemmateriaal aan leerlingen voor wie het niet specifiek is geconstrueerd. Daarom is er besloten dat het overlapmateriaal dat aan een bepaalde (half-)jaargroep kan worden voorgelegd alleen itemmateriaal mag bevatten dat specifiek voor die halfjaargroep is geconstrueerd en voor de twee belendende halfjaargroepen. Voor M5 betekent dit dat de leerlingen in het kalibratie- en normeringsonderzoek items voorgelegd krijgen die specifiek voor M5 zijn geconstrueerd, en een (minderheid aan) items die geconstrueerd zijn voor E4 en E5.

Het macrodesign is weergegeven in onderstaande figuur.

Figuur 4.1 Macrodesign LVS 3.0 Begrijpend lezen



De items die voor de overlap of verankering zorgen, duiden we in het macrodesign aan met ank, gevolgd door 2 cijfers. Zo duidt ank34 de groep items aan die enerzijds bestaat uit items geconstrueerd voor E3 en anderzijds uit items geconstrueerd voor M4. Die items zijn dus zowel eind groep 3 als medio groep 4 afgenomen. De groep items ank44 bevat items voor M4 en E4, die dus zowel medio groep 4 als eind groep 4 zijn afgenomen. Een item kan hoogstens in één (overlap)groep voorkomen, dat wil zeggen:

de ank-blokjes hebben geen gemeenschappelijke items met elkaar en ook niet met de reguliere blokjes E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8.

Longitudinale opzet

Een volledig longitudinaal design impliceert dat een cohort leerlingen gevolgd wordt van E3 tot en met M8. Een dergelijk design heeft een aantal zwaarwegende nadelen. Het is onvermijdelijk dat er uitval plaats zal vinden. Bij ernstige selectieve uitval wordt het steeds ingewikkelder om betrouwbare normen op te stellen. Bovendien is een longitudinale studie belastend voor de deelnemende scholen en leerlingen. Dit brengt het risico mee van ongewenste en moeilijk controleerbare neveneffecten. Daarom is ervoor gekozen het longitudinale karakter van het onderzoek in te perken, en aan de deelnemende scholen te vragen deel te nemen op drie opeenvolgende meetmomenten, waarbij het startmoment verspreid is voor verschillende scholen. Bijvoorbeeld: school A start met groep 4 op het eindmoment van schooljaar x en zal eveneens deelnemen aan de opvolgende momenten M5 (schooljaar x+1) en E5 (schooljaar x+1). School B zal starten op moment M5 (schooljaar y) en zal eveneens deelnemen aan de opvolgende momenten E5 (schooljaar y) en M6 (schooljaar y+1). Op deze manier wordt rekening gehouden met de belasting voor scholen en worden toch de benodigde longitudinale data verkregen.

Aansluitend bij de verbondenheid van het design via opeenvolgende toetsmomenten en de longitudinale opzet zal de kalibratie per leerjaar worden uitgevoerd op een beperkt deel van de gemeenschappelijke schaal. De kalibratie zal plaatsvinden op basis van de verzamelde data voor dat leerjaar op de afname-momenten, aangevuld met de gegevens van het voorgaande en het opvolgende afnamemoment. In het geval van leerjaar 4 met afnamemomenten M4 en E4 vindt de kalibratie plaats op basis van afname-momenten E3, M4, E4 en M5. Dit sluit aan bij de inhoudelijke kenmerken van de aangeboden opgaven, een sterke leerling in groep 4 zal wel opgaven uit groep 5 kunnen maken, maar geen opgaven uit groep 8 omdat deze qua inhoud nog niet allemaal zijn behandeld. Op deze manier kan dus beter rekening gehouden worden met de uitbreidingen in het onderwijsaanbod. Voor kalibratie en normering van de toetsen van elke jaargroep zal op een gedeelte van het eerder vermelde design worden gefocust. In het geval van groep 4 betreft het dus de regels 1 tot en met 3 uit het macrodesign (figuur 4.1). Dit gedeelte uit het macrodesign wordt hieronder weergegeven in figuur 4.2.

Figuur 4.2 Gedeelte macrodesign waarop kalibratie leerjaar 4 is gebaseerd'

Januari 2013		ank34	M4	ank44				
Juni 2013	E3	ank34		ank44	E4	ank45		
Januari 2014						ank45	M5	ank55

Opgemerkt dient te worden dat de normering onafhankelijk is van de aangeboden items, mits deze qua inhoud passen bij de jaargroep en passen op de kalibratieschaal. De normering wordt immers gebaseerd op de vaardigheid op dat afnamemoment. De afgenomen toets is slechts een middel om de vaardigheid te bepalen. De opzet van de kalibratie en de normering zullen in de volgende paragrafen verder worden beschreven.

Om de prestaties van een leerlingen en groepen te kunnen blijven volgen, zullen deze op een overkoepelende schaal worden geplaatst door gebruik te maken van een transformatie. Deze transformatie wordt afgeleid uit de overlappende populaties op de kalibratie en normering per jaargroep.

De overlappende jaargroepen op opvolgende schalen bestaan uit dezelfde leerlingen in beide kalibraties en hebben per definitie dezelfde vaardigheidsverdeling. Om deze reden kan uit de vaardigheidsverdelingen van die jaargroepen de transformatie berekend worden.

4.2 De kalibratie

In hoofdstuk 2 zijn in algemene zin de procedures beschreven die leiden tot gekalibreerde opgavenbanken. Tevens gaat dat hoofdstuk in op het meetmodel dat ten grondslag ligt aan de toetsen Begrijpend lezen. In deze paragraaf gaan we nog wat gedetailleerder in op het kalibratieonderzoek. Eerst komt de opzet daarvan aan de orde (paragraaf 4.2.1) en beschrijven we de stappen die in het kader van de kalibratie zijn gezet (paragraaf 4.2.2). In paragraaf 4.2.3 geven we resultaten van analyses die duidelijk maken dat de kalibratie geslaagd genoemd kan worden.

4.2.1 De opzet van de kalibratie

Prestaties van leerlingen blijken al snel na publicatie van een toets te verschuiven, omdat bij het onderzoek dat ten grondslag ligt aan de normering sprake is van low stakes afnamesituaties (Keuning et al. 2015). Bij de ontwikkeling van LVS 3.0 is geprobeerd om bias in de normen te vermijden door de afnamesituatie waarin de toets wordt afgenomen zoveel mogelijk te laten lijken op de situatie na uitgave. Er is gekozen voor een *embedded field* onderzoek, waarin nieuw ontwikkelde items voor de derde generatie van het Cito Leerlingvolgsysteem primair en speciaal onderwijs (LVS 3.0) meelopen in de al bestaande en op scholen toegepaste toetscyclus. Aan de reguliere afname van de toetsen Begrijpend lezen uit de tweede generatie van het Cito Leerlingvolgsysteem primair en speciaal onderwijs (LVS 2.0) zijn eenmalig twee taken met nieuw materiaal toegevoegd. In totaal lieten scholen hun leerlingen drie taken maken; een taak uit LVS 2.0 en twee taken met nieuw materiaal voor de derde generatie. Bij de leerlingen was onbekend welke taken de nieuwe opgaven bevatten. Tevens was voor de leerlingen onbekend dat de gegevens ook voor onderzoeksdoeleinden werden gebruikt. Voor deze opzet werd gekozen opdat motivatie-effecten de verzamelde gegevens voor het normeringsonderzoek zo min mogelijk zouden beïnvloeden. Een belangrijk tweede voordeel van deze aanpak is dat de normeringssteekproef aangevuld kan worden met resultaten uit dataretour van de tweede generatie LVS-toetsen (zie Keuning et al., 2015).

In figuur 4.3 en 4.4 worden de *embedded field* designs weergegeven voor respectievelijk toetsversie M4 en toetsversie E4. In het design kunnen we zien dat er negen toetsversies M4 en negen toetsversies E4 zijn afgenomen. Elke leerling maakte volgens het design een taak uit de M4- dan wel E4-toets LVS 2.0. Daarnaast maakte elke leerling twee taken van de beoogde uitgave LVS 3.0. Om de opzet van het design goed te kunnen doorgronden, is het nodig om zich te realiseren dat in LVS 3.0 ook tussentoetsen worden uitgegeven. Deze tussentoetsen vormen qua niveau een tussenversie tussen twee toetsen voor opvolgende afnamemomenten en worden verantwoord tezamen met het hogere afnamemoment. Zo kan de tussentoets E4M5 gebruikt worden voor leerlingen op het moment E4 voor wie verwacht wordt dat de E4-toets te makkelijk zal zijn, maar eveneens voor de leerlingen op het moment M5 voor wie verwacht wordt dat de M5-toets te moeilijk zal zijn. De verantwoording van de tussentoets E4M5 blijft hier verder buiten beschouwing en zal worden behandeld in de verantwoording voor de toetsen bestemd voor groep 5. De taken E3M4 deel 1 en E3M4 deel 2 vormen tezamen de beoogde uitgave E3M4 LVS 3.0 en zijn ook eind groep 3 afgenomen. De taken M4 deel 1 en M4 deel 2 vormen tezamen de beoogde uitgave M4 LVS 3.0. De taken M4E4 deel 1 en M4E4 deel 2 vormen tezamen de beoogde uitgave M4E4 LVS 3.0. De taken E4 deel 1 en E4 deel 2 vormen tezamen de beoogde uitgave E4 LVS 3.0. Tot slot vormen de taken E4M5 deel 1 en E4M5 deel 2 de beoogde uitgave E4M5 LVS 3.0; deze zullen ook medio groep 5 worden afgenomen. Voor zowel E3M4, M4, M4E4, E4 als E4M5 is een 'reservetaak' meegenomen in het normeringsonderzoek. Deze opgaven zouden in de uiteindelijke uitgave alleen worden gebruikt indien er onverwachte problemen naar voren kwamen met betrekking tot de beoogde taken voor de uitgave E3M4, M4, M4E4, E4 en E4M5. Zoals te zien in het designs vormden de delen M4 LVS 2.0 en E4 LVS 2.0 een stevig anker tussen de toetsboekjes voor respectievelijk afnamemoment M4 en afnamemoment E4 en werd er tussen de beoogde reguliere toetsen en aangrenzende tussentoetsen voor LVS 3.0 geankerd in de normeringsonderzoeken M4 LVS 3.0 en E4 LVS 3.0. Omdat de tussentoetsen ook werden meegenomen in twee opvolgende normeringsonderzoeken is er ook een ankering over de toetsmomenten heen.

De leerlingen maakten zowel oud als nieuw materiaal. Door deze opzet kan de zogenoemde dataretour LVS 2.0 worden meegenomen in het vaststellen van de normering voor de uitgave LVS 3.0 groep 4. Ook kunnen we door deze opzet de normering van de nieuw uit te geven toetsen vergelijken met de normering van LVS 2.0 en kan de continuïteit tussen LVS 2.0 en LVS 3.0 in beeld worden gebracht (zie hoofdstuk 6 over validiteit).

Figuur 4.3 Design LVS 3.0 M4

Toets	M4 LVS 2.0	E3M4 deel 1	E3M4 deel 2	E3M4 reserve	M4 deel 1	M4 deel 2	M4 reserve	M4E4 deel 1	M4E4 deel 2	M4E4 reserve	Leerlingen
1											151
2											206
3											202
4											198
5											203
6											175
7											189
8											176
9											173

In het normeringsonderzoek M4 zijn 235 items voorgelegd aan 1673 leerlingen medio groep 4 verdeeld over de 9 boekjes zoals aangegeven in de laatste kolom van figuur 4.3. Elk boekje bestond uit 69 tot 75 opgaven verdeeld over 3 taken. De 25 opgaven in de LVS 2.0 taak werden door alle 1673 leerlingen gemaakt. De overige opgaven kwamen in twee boekjes voor en werden gemiddeld door 372 leerlingen gemaakt. Bij alle opgaven werd voldaan aan de eis dat deze minimaal bij 150 leerlingen moest zijn afgenomen.

Figuur 4.4 Design LVS 3.0 E4

Toets	E4 LVS 2.0	M4E4 deel 1	M4E4 deel 2	M4E4 reserve	E4 deel 1	E4 deel 2	E4 reserve	E4M5 deel 1	E4M5 deel 2	E4M5 reserve	Leerlingen
1											280
2											256
3											264
4											259
5											273
6											278
7											245
8											281
9											285

In het normeringsonderzoek E4 zijn 244 items voorgelegd aan 2421 leerlingen eind groep 4 verdeeld over de 9 boekjes zoals aangegeven in de laatste kolom van figuur 4.3. Elk boekje bestond uit 69 tot 75 opgaven verdeeld over 3 taken. De 25 opgaven in de LVS 2.0 taak werden door alle 2421 leerlingen gemaakt. De overige opgaven kwamen in twee boekjes voor en werden gemiddeld door 538 leerlingen gemaakt. Bij alle opgaven werd voldaan aan de eis dat deze minimaal bij 150 leerlingen moest zijn afgenomen. Op grond van de gegevens uit de kalibratie van de normeringsonderzoeken is de definitieve selectie van items gemaakt voor de uitgave van de LVS 3.0 toetsen E3M4, M4, M4E4 en E4.

4.2.2 De stappen in de kalibratie

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

De statistische toetsen in OPLM hebben goede statistische en asymptotische eigenschappen, daar OPLM behoort tot de exponentiële familie, met de gewogen somscore:

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

als een 'afdoende statistiek' (*sufficient statistic*) voor de vaardigheid θ . Dit betekent dat alle informatie in de data met betrekking tot de vaardigheid in deze statistiek aanwezig is. Hiervan wordt gebruikgemaakt bij de statistische toetsen in OPLM. Het basisprincipe van de statistische toetsen in OPLM is dat op grond van de afdoende statistiek s de personen in de data kunnen worden gegroepeerd. En binnen deze groepen kan de verwachte proportie goede antwoorden op een item onder het model, $p(+|s)$, vergeleken worden met de feitelijk geobserveerde proportie goede antwoorden, $prop(+|s)$. Via de basisvergelijking van OPLM kunnen we eenvoudig de conditionele kans op het goed beantwoorden van de items afleiden en daarmee kunnen we $p(+|s)$ evalueren, $prop(+|s)$ volgt uit de data. Discrepancies tussen $p(+|s)$ en $prop(+|s)$ duiden op schendingen van het model. Deze discrepanties vormen de basis voor de diverse statistische toetsen in OPLM. De toetsingsgrootte voor de veronderstelde discriminatie-indices is gegeven door:

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

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

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

Deze zogeheten S-toets heeft een χ^2 verdeling onder het model. Als globale modeltoets is de R1c-toets (Glas, 1988) geschikt. Ook de distributie van de rechteroverschrijdingskansen van alle afzonderlijke S-toetsen komt hiervoor in aanmerking. Als we deze S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn op het (0,1) interval. Kortom, als we afzien van de formeel-statistische achtergrond van de gehanteerde toetsen, kan de kalibratieprocedure als volgt worden samengevat:

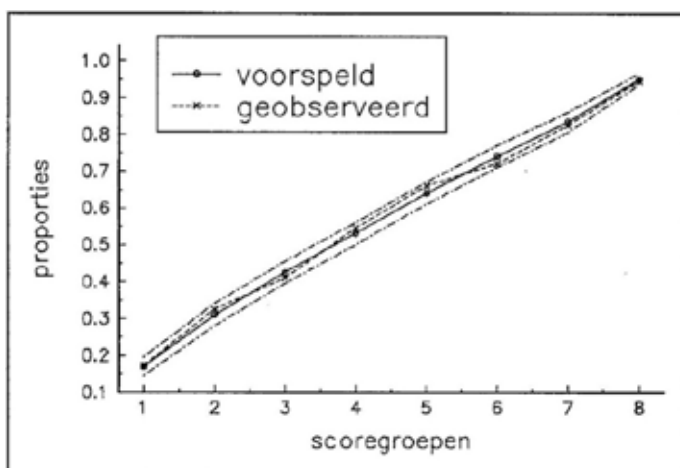
- 1 Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
- 2 Vervolgens schatten we de itemparameters met behulp van de CML-methode.
- 3 Met behulp van de M-toetsen controleren we of de discriminatie-indices goed zijn ingesteld.
- 4 Een volgende controle betreft de overschrijdingskansen van de S-toetsen en een grafische modelcontrole door middel van het programma WOPPLOT (grafische inspectie van de ICC's).
- 5 Vervolgens vindt een globale modelcontrole plaats in de vorm van een R1c-toets en de verdeling van de overschrijdingskansen van de S-toetsen.

De stappen 1 tot en met 5 worden een aantal malen doorlopen tot het resultaat bevredigend is. Afhankelijk van de uitkomsten kunnen items worden verwijderd. Ook inhoudelijke overwegingen spelen een rol in dit beslissingsproces (zie hiervoor hoofdstuk 2 over de achtergronden van de toetsinhoud).

4.2.3 Toetsing van het IRT-model

Het is niet eenvoudig om de kwaliteit van de kalibratie aan te tonen. De belangrijkste statistische instrumenten om de passing van een opgave in het IRT-model te bewerkstelligen en uiteindelijk te documenteren betreffen de hierboven al besproken S-toetsen. Het lastige daarvan is, dat de toetsing voor een groot deel visueel gebeurt. Dit kunnen we illustreren aan de hand van figuur 4.5 (zie Staphorsius, 1994, blz. 239). Figuur 4.5 beeldt voor een opgave de gegevens af waarop de betreffende S-toetsen gebaseerd zijn (zie handleiding OPLM: Verhelst; 1992). Ten behoeve van deze toetsing wordt de totale groep van leerlingen die een verzameling opgaven gemaakt heeft, ingedeeld in een aantal (meestal acht) scoregroepen. Elke groep bestaat uit leerlingen met een ongeveer even hoge score. De geobserveerde proporties juiste antwoorden van deze groepen (telkens gesymboliseerd door een x) zijn door de middelste stippelijijn verbonden. De volle lijn daarentegen verbindt de proporties die op grond van de parameterschattingen voorspeld kunnen worden. De twee buitenste lijnen geven het 95%-betrouwbaarheidsinterval aan. De breedte van dit interval is in belangrijke mate afhankelijk van het aantal leerlingen dat de opgave heeft beantwoord. Uit de figuur blijkt heel duidelijk dat de geobserveerde proporties, zoals bedoeld, binnen het 95%- betrouwbaarheidsinterval van de (geschatte) voorspelde proporties liggen, en dit komt in grote lijnen overeen met een niet-significante S-toetsingsgrootte (Verhelst, et al., 1994).

Figuur 4.5 Grafische voorstelling van een S_i -toets

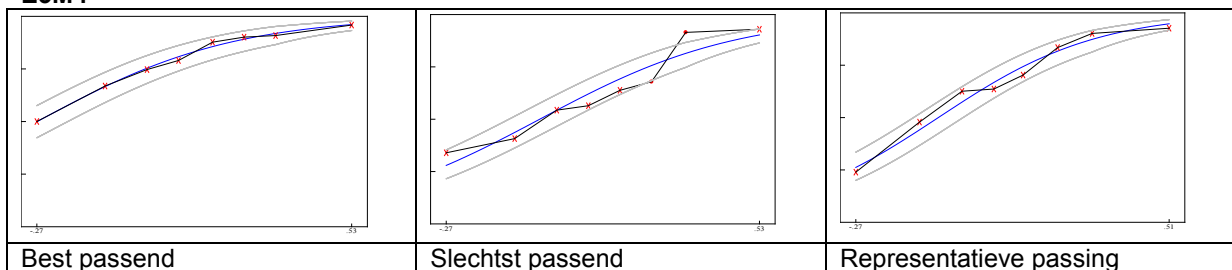


Het is ondoenlijk om voor alle opgaven dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons steeds per toetsversie tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.6 illustreren dat voor de toetsen E3M4, M4, M4E4 en E4 zelfs bij de

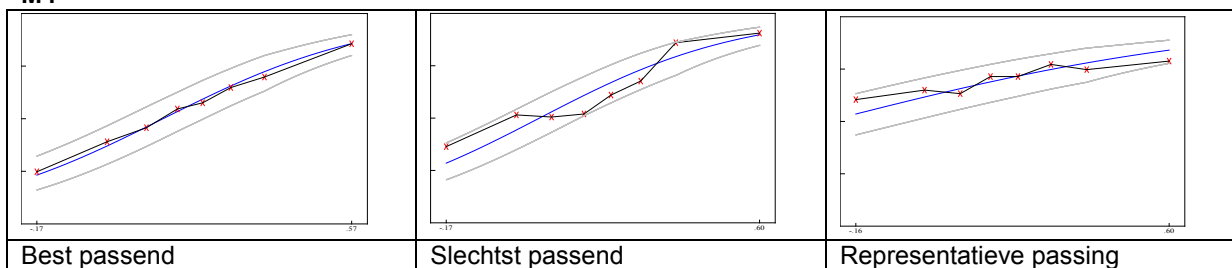
slechtst passende opgave sprake is van een zeer aanvaardbaar beeld. Er wordt in dit geval voor een deel (van de onderscheiden scoregroepen) niet beantwoord aan de eis dat de geobserveerde proportie binnen het 95%- betrouwbaarheidsinterval van de geschatte proporties ligt. Dit beeld doet zich slechts bij enkele opgaven voor die dan ook een uitzondering vormen. De overige opgaven voldoen voor alle scoregroepen wel aan die eis. De afbeeldingen voor de representatieve en best passende opgaven illustreren dit. Dit leidt tot de conclusie dat bij vrijwel alle opgaven in de Begrijpend lezen toetsen een grafische voorstelling van de S-toetsing hoort die in grote lijnen met figuur 4.5 overeenkomt. Dit is, zeker gezien de relatief grote aantallen observaties die in het geding zijn, een zeer sterke aanwijzing dat het meetinstrument en het meetmodel dat ontwikkeld is, respectievelijk gebruikt is, adequaat zijn om het gedrag van de leerlingen te verklaren. Bovendien blijkt, en dat is vanuit theoretisch oogpunt nog belangrijker, dat gemeten verschillen in gedrag tussen de leerlingen te verklaren zijn door één unidimensioneel concept.

Figuur 4.6 Voorbeelden van S-toetsen voor de Toets LVS 3.0 Begrijpend lezen E3M4, M4, M4E4 en E4 met de best passende, de slechtst passende en een qua passing representatieve opgave

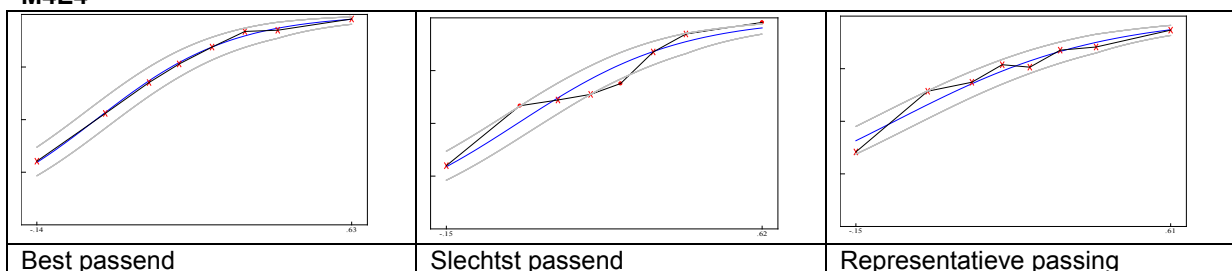
E3M4



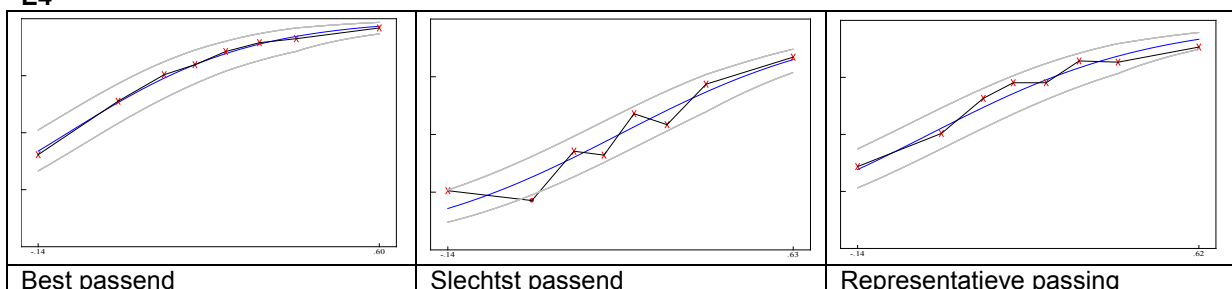
M4



M4E4



E4



In feite kan men bij de kalibratie beter varen op deze grafische weergaven dan op toetsingsresultaten in termen van exacte getallen en de significantie daarvan. Niettemin zijn er bij de kalibratie S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij zijn we vooral geïnteresseerd in de distributie van de overschrijdingskansen van deze verzameling toetsingsresultaten. Tabel 4.1 waarin het (0,1) interval is opgedeeld in tien gelijke stukken, geeft een beeld van de uitkomsten bij een kalibratie van alle opgaven van de toetsen LVS 3.0 Begrijpend lezen E3M4, M4, M4E4 en E4. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan .01, respectievelijk .05. Het is duidelijk dat voor de toetsen de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Dit resultaat geeft een bevestiging van het eerder geschetste beeld, dat met uitzondering van enkele opgaven, sprake is van niet-significante S-toetsen. Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensioneel construct representeren.

Tabel 4.1 Verdeling van overschrijdingskansen bij S-toetsen voor toetsen E3M4, M4, M4E4 en E4

	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
E3M4	6	2	5	1	4	5	3	2	3	3	3
M4	0	2	2	6	3	5	5	5	5	0	4
M4E4	2	7	2	4	6	2	2	1	5	6	1
E4	0	4	2	3	7	6	2	1	2	2	6

In tabel 4.2 zijn de R1c-waarden weergegeven voor dezelfde afnames waarvoor in tabel 4.1 de resultaten van de S-toetsen zijn weergegeven. R1c is een statistiek die zicht geeft op de modelpassing van de toets als geheel. Voor een acceptabele modelfit geldt als vuistregel dat R1c bij voorkeur niet significant zou moeten zijn en niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden (df). De modelpassing van de toetsen voldoet aan deze voorwaarden. Voor alle toetsen E3M4, M4, M4E4 en E4 geldt dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootte is significant voor de toetsen E3M4, M4, M4E4 en E4. Aan dit laatste moet bij steekproeven met dergelijke omvang niet te veel waarde worden gehecht.

Tabel 4.2 R1c-waarden voor E3M4, M4, M4E4 en E4

Toetsversie	R1c	df	p
E3M4	1097.163	777	<.005
M4	395.706	286	<.005
M4E4	376.120	308	<.005
E4	339.866	278	<.010

Ten slotte bespreken we nog een methode om de modelpassing te verantwoorden die wordt besproken in het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer & Sijtsma, 2010, p. 40). Het betreft hier een poging om de nauwkeurigheid van de itemparameterschattingen te beoordelen op basis van een constante (in het COTAN-Beoordelingssysteem met 'c' aangeduid) die weergeeft hoe de relatie is tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie. Het beoordelingssysteem geeft ook richtlijnen voor het beoordelen van de grootte van deze 'c'. Deze dient te worden beoordeeld als goed als de waarde lager is dan of gelijk aan .20. Waarden tussen .30 en .40 kunnen nog als voldoende worden beschouwd. In tabel 4.3 zijn gemiddelde en range van deze waarden voor alle items in de toetsen weergegeven. De gemiddelde waarde van de constante is uitstekend te noemen. Voor geen enkele opgave is c groter dan .20.

Tabel 4.3 *Nauwkeurigheid van de itemparameterschattingen (constante 'c')*

Toetsmoment	Constante 'c'	
	Range	Gemiddelde
E3M4	.043-.104	.063
M4	.061-.169	.095
M4E4	.042-.148	.069
E4	.055-.136	.078

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen LVS 3.0 Begrijpend lezen E3M4, M4, M4E4 en E4 de kalibratie geslaagd is. Hiermee is het laatste woord nog niet gezegd over de validiteit, maar het kalibratieonderzoek brengt in ieder geval een essentieel aspect van het validiteitsvraagstuk naar voren: de rechtvaardiging van wat in de meeste toetstoepassingen gebruikelijk is, namelijk het reduceren van alles wat de leerling heeft geantwoord tot een enkele toetsscore (of afgeleid daarvan, een enkele schatting van zijn onderliggende vaardigheid). De kalibratie-analyse, als puur formeel proces, kan geen uitspraken doen over de inhoudsvaliditeit of over de constructvaliditeit als antwoord op de vraag: hoe kan worden aangetoond dat het concept dat de items in de bank meten dekkend is voor en samenvalt met het construct dat we in de toetsen LVS 3.0 Begrijpend lezen proberen te meten (zoals dat in het didactisch en het wetenschappelijk forum wordt bedoeld)? In hoofdstuk 6 over validiteit zal worden nagegaan of de gemeten concepten inderdaad overeenkomen met het begrip zoals bedoeld. De vraag is dan in het geval van het onderdeel Begrijpend lezen: kan het unidimensionele concept onder de opgaven in de opgavenbank Begrijpend lezen inderdaad worden opgevat als de vaardigheid 'begrijpend lezen'? Een geslaagde kalibratie op een unidimensioneel construct beschouwen we als een noodzakelijke voorwaarde voor deze begripsvaliditeit.

4.3 De normering

Sinds schooljaar 2013/2014 wordt door Cito een nieuwe werkwijze voor het normeren van leerling-olgsysteemtoetsen toegepast. Deze werkwijze wordt gebruikt bij het monitoren van de normering van inmiddels uitgegeven toetsen, maar wordt ook gebruikt bij de normering van de nieuw uit te geven toetsen, zo ook bij de derde generatie toetsen voor Begrijpend lezen. De werkwijze die we hieronder beschrijven, komt uit Keuning et al. (2015). Allereerst besteden we aandacht aan de opzet van het normeringsonderzoek, de gehanteerde procedures en de aantallen leerlingen per afnamemoment (paragraaf 4.3.1). Vervolgens komt in paragraaf 4.3.2 de representativiteit van de normsteekproeven aan de orde. De paragraaf wordt afgerond met een presentatie van de resultaten van de normering (i.e. de kenmerken van de vaardigheidsverdelingen op de onderscheiden afnamemomenten; paragraaf 4.3.3).

4.3.1 Opzet

Tijdens het *embedded field* normeringsonderzoek zoals omschreven in paragraaf 4.2.1 worden data verzameld. Om deelnemers te werven voor het normeringsonderzoek zijn scholen aangeschreven. Voor het *embedded field* normeringsonderzoek is een representatieve steekproef getrokken uit de verzameling van alle basisscholen in Nederland. Dit is gedaan vanuit het bij Cito gebruikelijke steekproefkader dat bepaald wordt door regio, urbanisatiegraad en schooltype (zie verderop voor een omschrijving van deze achtergrondvariabelen). De dekking van de LVS toetsen Begrijpend lezen 2^e generatie is bijzonder hoog: de toetsen worden door 85% tot 90% van de scholen toegepast. Voor de deelnemende scholen aan het normeringsonderzoek is nagegaan of zij als groep afweken van wat men voor de totale populatie van scholen zou mogen verwachten. De gemiddelde score op de Cito Eindtoets Basisonderwijs bleek voor de scholen in het normeringsonderzoek niet af te wijken van het populatiegemiddelde voor deze toets. Dit is voor meerdere jaren onderzocht.

Scholen werd gevraagd om op drie opeenvolgende momenten deel te nemen. Zoals in het design zichtbaar is (zie figuur 4.1), was normeringsonderzoek M4 het eerste afnamemoment. Hiervoor zijn geen herhalingscholen benaderd. Van de deelnemende scholen op afnamemoment M4 zijn enkele van deze scholen benaderd als herhalingscholen voor het normeringsonderzoek op afnamemoment E4. De normeringsgroep van afnamemoment E4 bestond dus deels uit herhalingscholen.

Voor het totaal aantal scholen in Nederland (7168 scholen) is een indeling gemaakt naar LOVS strata (0 t/m 10 procent, 11 t/m 25 procent, 26 t/m 40 procent en 41 procent of meer gewichtsl leerlingen) bij schoolgrootte (meer dan 200 leerlingen dan wel minder dan 200 leerlingen). Dit resulteerde in 8 groepen. Niet over alle scholen waren deze gegevens bekend. De scholen waarvan deze gegevens onbekend waren zijn ingedeeld in een restcategorie. Vervolgens zijn clustersteekproeven getrokken op dusdanige wijze dat de 9 groepen representatief waren vertegenwoordigd in de steekproef. Bij de steekproeftrekking werd de inschatting van deelnamebereidheid gebaseerd op voorgaande wervingen. Uit deze wervingen bleek dat het gemiddelde aantal deelnemende leerlingen per school 24 was en de deelnamebereidheid aan normeringsonderzoeken 6%.

Voor het normeringsonderzoek M4 zijn uiteindelijk 1221 scholen aangeschreven. Omdat na de eerste aanschrijvingsronde slechts 5% van de scholen uit de steekproef bereid bleek deel te nemen aan het normeringsonderzoek zijn in een tweede wervingsronde 340 extra scholen aangeschreven. Uiteindelijk resulteerde dit in een deelnamebereidheid van zo'n 5% van de overige aangeschreven scholen. In totaal meldden zich 75 scholen aan voor het normeringsonderzoek met in totaal 1826 leerlingen. Uiteindelijk was het aantal scholen dat daadwerkelijk gegevens aanleverde gelijk aan 73 en het aantal leerlingen aan 1708. Van de 1708 leerlingen die hebben deelgenomen aan het normeringsonderzoek konden 1673 leerlingen van 73 verschillende basisscholen worden meegenomen in de data-analyses.

Voor het normeringsonderzoek E4 zijn uiteindelijk in totaal 77 herhalingscholen (scholen die aangegeven hadden deel te willen nemen) aangeschreven en 854 extra scholen. Omdat na de eerste aanschrijvingsronde 53% van de herhalingscholen uit het normeringsonderzoek M4 ook bereid bleek deel te nemen aan het normeringsonderzoek E4 en minder dan 3% van de scholen uit de aanvullende steekproef bereid bleek deel te nemen aan het normeringsonderzoek zijn in een tweede wervingsronde 1400 extra scholen aangeschreven. Uiteindelijk resulteerde dit in een deelnamebereidheid van zo'n 53% van de herhalingscholen en 3% van de overige aangeschreven scholen. In totaal meldden zich 108 scholen aan voor het normeringsonderzoek met in totaal 2657 leerlingen. Uiteindelijk was het aantal scholen dat daadwerkelijk gegevens aanleverde gelijk aan 104 en het aantal leerlingen aan 2498. Van de 2498 leerlingen die hebben deelgenomen aan het normeringsonderzoek konden 2417 leerlingen van 103 verschillende basisscholen worden meegenomen in de data-analyses.

Voor het bepalen van de normering worden de gegevens aangevuld met gegevens uit Cito dataretour. Vanzelfsprekend worden de data die via Cito dataretour binnenkomen opgeschoond voordat ze gebruikt worden. Uit de bestanden worden de volgende categorieën leerlingen verwijderd:

- Leerlingen uit het speciaal onderwijs en leerlingen voor wie het onderwijstype onbekend is.
- Leerlingen van scholen die het LVS selectief inzetten. In de hogere leerjaren blijken sommige scholen het LVS namelijk alleen in te zetten bij zwakkere leerlingen (zie Keuning, 2011).

- Leerlingen die op hetzelfde afnamemoment meerdere toetsen van dezelfde vaardigheid maken. Alleen de gegevens van de toets die bij het afnamemoment hoort, worden behouden. Daarnaast worden de scholen verwijderd die ook aan de *embedded field* normeringsonderzoeken deelnemen.

Er is voor gekozen om alleen data te selecteren van het schooljaar waarin ook het normeringsonderzoek heeft plaatsgevonden. Er wordt naar gestreefd om de uiteindelijke normeringssteekproef voor ongeveer 50 procent te baseren op gegevens uit het *embedded field* normeringsonderzoek en voor 50 procent op gegevens uit Cito dataretour. De streefverhouding kan desgewenst ook anders gekozen worden, maar het ligt niet voor de hand om het aandeel van het ene gegevensbestand veel groter te maken dan het aandeel van het andere gegevensbestand. Door Cito dataretour een groter gewicht te geven, neemt het percentage leerlingen dat de nieuwe LVS 3.0 toetsen maakt namelijk verhoudingsgewijs af. Met het oog op de constructie en validering van LVS 3.0 is dit onwenselijk. Door het *embedded field* normeringsonderzoek een groter gewicht te geven, neemt de hoeveelheid data die volledig in de feitelijke toetssituatie verzameld zijn af. Dit is een gemiste kans. Juist het combineren van het *embedded field* normeringsonderzoek met Cito dataretour biedt grote voordelen ten opzichte van alternatieve onderzoeksdesigns. Enerzijds wordt er op deze manier voor gezorgd dat de toetsresultaten die gebruikt worden bij het bepalen van de normen zoveel mogelijk in de feitelijke toetssituatie verzameld zijn. Anderzijds is het mogelijk om via Cito dataretour de “kwaliteit” van het *embedded field* normeringsonderzoek te checken. Een belangrijke randvoorwaarde is wel dat de uiteindelijke normeringsteekproef representatief is voor de landelijke populatie van scholen en leerlingen. Representativiteit van de normeringssteekproef zoals die samengesteld wordt op basis het *embedded field* normeringsonderzoek (± 50 procent) en Cito dataretour (± 50 procent) is te realiseren door bij de selectie van data uit Cito dataretour rekening te houden met relevante achtergrondvariabelen. Bij de normering van LVS 3.0 wordt rekening gehouden met de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *sekse*. De verschillende variabelen zijn als volgt gedefinieerd:

- **Regio.** Bij de definitie van de variabele *regio* is uitgegaan van de CBS-indeling naar landsdeel. Dit betekent dat er vier regio's onderscheiden zijn. Regio *noord* omvat de provincies Groningen, Friesland en Drenthe; regio *oost* de provincies Overijssel, Gelderland en Flevoland; regio *west* de provincies Utrecht, Noord-Holland, Zuid-Holland en Zeeland en regio *zuid* de provincies Noord-Brabant en Limburg.
- **Urbanisatiegraad.** Bij de definitie van de variabele *urbanisatiegraad* is er voor gekozen om de indeling naar vijf niveaus die gebruikelijk is bij het CBS te reduceren tot een tweedeling in enerzijds niet tot matig verstedelijkt (platteland) en anderzijds sterk tot zeer sterk verstedelijkt (stad). Een dergelijke tweedeling blijkt in de praktijk goed te volstaan (cf. Van Boxtel & Hemker, 2009).
- **Schooltype.** Bij de definitie van de variabele *schooltype* is gebruikgemaakt van de formatiegewichten van de leerlingen binnen een school volgens de meest recente regeling van OCW. Daarin worden drie niveaus onderscheiden die gebaseerd zijn op het opleidingsniveau van de ouders:
 - 0.0 één van de ouders of beide ouders heeft of hebben een opleiding gehad uit categorie 3
 - 0.3 beide ouders of de ouder die belast is met de dagelijkse verzorging heeft of hebben een opleiding uit categorie 2 gehad
 - 1.2 één van de ouders heeft een opleiding gehad uit categorie 1 en de ander een opleiding uit categorie 1 óf 2

In deze indeling wordt verwezen naar de volgende categorieën in het opleidingsniveau van de ouders: 1 = maximaal basisonderwijs of (V)SO-ZMLK, 2 = maximaal LBO/VBO, praktijkonderwijs of VMBO basis- of kaderberoepsgerichte leerweg, en 3 = overig VO en hoger. Leerlingen met een formatiegewicht van 0.3 of 1.2 zijn te definiëren als achterstandsl leerlingen. Scholen zijn ingedeeld naar het percentage achterstandsl leerlingen volgens een indeling in vier typen: (1) percentage achterstandsl leerlingen [0, .10), (2) percentage achterstandsl leerlingen [.10, .25), (3) percentage achterstandsl leerlingen [.25, .40) en (4) percentage achterstandsl leerlingen [.40, 1].

- **Sekse.** Bij de variabele *sekse* is een tweedeling naar jongens en meisjes gehanteerd.

Het is niet mogelijk om expliciet rekening te houden met de variabele *etniciteit*, omdat (a) er geen eenduidige referentiegegevens voor de populatie bekend zijn, en (b) Cito dataretour weinig tot geen informatie bevat over de etnische herkomst van leerlingen. Onderzoek heeft echter laten zien dat de verdeling naar etnische herkomst sterk samenhangt met de verdeling naar urbanisatiegraad en schooltype (Hemker, Kordes en Van Weerden, 2011). Om deze reden is aangenomen dat de uiteindelijke normeringsteekproef voldoende representatief is naar etnische herkomst als de verdeling naar urbanisatiegraad en schooltype overeenkomt met de verdeling in de landelijke populatie.

Bij het selecteren van data uit Cito dataretour wordt rekening gehouden met vier achtergrondvariabelen die samen $4 \times 2 \times 4 \times 2 = 64$ verschillende categorieën representeren. De variabelen *regio*, *urbanisatiegraad* en *schooltype* zijn op het niveau van de school gedefinieerd. De variabele *sekse* is op het niveau van de leerling gedefinieerd. Het is niet goed mogelijk om bij het selecteren van data tegelijkertijd rekening te houden met school- én leerlingvariabelen. Daarom vindt de dataselectie in twee stappen plaats. In de eerste stap worden iteratief scholen uit Cito dataretour toegevoegd aan de dataset met normeringsgegevens. Niet elke school heeft daarbij evenveel kans om geselecteerd te worden. Bij de selectie wordt namelijk rekening gehouden met de regio en de urbanisatiegraad van de school en het aantal achterstandsleerlingen. De kans w_{ijk} dat een school met regio i , urbanisatiegraad j en schooltype k geselecteerd wordt, hangt af van het reeds geselecteerde aantal leerlingen N_S , het gewenste aantal leerlingen N_T , en het beschikbare aantal leerlingen in Cito dataretour N_D :

$$w_{ijk} = \frac{(n_{T,ijk} - n_{S,ijk}) \div (N_T - N_S)}{n_{D,ijk} \div N_D} = \frac{N_D (n_{T,ijk} - n_{S,ijk})}{n_{D,ijk} (N_T - N_S)},$$

waarbij vereist is dat $n_{S,ijk} \leq n_{T,ijk}$. Zoals we kunnen zien, wordt het percentage leerlingen dat (nog) gewenst is voor een bepaalde categorie (in dit geval de populatie) gedeeld door het percentage leerlingen dat via Cito dataretour beschikbaar is voor opname in die categorie (in dit geval de steekproef).

In geval $n_{S,ijk} > n_{T,ijk}$ is de kans w_{ijk} die uit de formule volgt negatief en niet toe te passen. Dat kan in twee situaties gebeuren. Ten eerste kan een bepaalde categorie in het licht van de gekozen N_T en de via de landelijke gegevens van DUO en/of CBS te bepalen $n_{T,ijk}$ oververtegenwoordigd zijn in de dataset met normeringsgegevens. In dat geval kan het selectiealgoritme niet gestart worden. De oplossing is om enkele scholen te verwijderen totdat voor alle categorieën geldt dat $n_{S,ijk} \leq n_{T,ijk}$. Ten tweede kan tijdens de selectie blijken dat een categorie oververtegenwoordigd raakt als we een bepaalde school vanuit Cito dataretour toevoegen aan de dataset met normeringsgegevens. Dit risico wordt groter naarmate het reeds geselecteerde aantal leerlingen N_S dichterbij het gewenste aantal leerlingen N_T komt te liggen. De oplossing is om N_T bij de berekening van de gewichten te vermenigvuldigen met een vrij te kiezen constante C en het algoritme te beëindigen in de eerste iteratie waarbij geldt dat $N_S \geq N_T$. Als constante C groot gekozen wordt, heeft het selectiealgoritme veel ruimte om scholen te kiezen. Het voordeel is dat het selectiealgoritme snel voorziet in een oplossing. Het nadeel is dat de verdeling naar *regio*, *urbanisatiegraad* en *schooltype* zoals we die na toepassing van het selectiealgoritme observeren in de normeringsteekproef nogal kan afwijken van de verdeling zoals we die wensen op basis van de landelijke gegevens van DUO en/of CBS. Als constante C klein gekozen wordt, zal het selectiealgoritme minder snel een oplossing vinden. Het eindresultaat zal doorgaans wel een grotere gelijkenis vertonen met de landelijke gegevens van DUO en/of CBS.

q Tot nu toe is bij de selectie van data uitsluitend rekening gehouden met de schoolvariabelen *regio*, *urbanisatiegraad* en *schooltype*. De leerlingvariabele *sekse* is nog niet in beschouwing genomen. Dat gebeurt in de tweede stap. Als blijkt dat de normeringsteekproef die is samengesteld in de eerste stap niet representatief is met betrekking tot de variabele *sekse*, dan wordt een tweede steekproeftrekking uitgevoerd. Eerst wordt op basis van de landelijke gegevens van CBS en de geobserveerde aantallen in de normeringsteekproef de kans w_q bepaald dat een leerling met sekse q in een representatieve

normeringssteekproef zit:

$$w_q = \frac{n_{T,q} \div N_T}{n_{S,q} \div N_S} = \frac{n_{T,q} N_S}{N_T n_{S,q}}.$$

Zoals we kunnen zien, wordt het gewenste percentage leerlingen in categorie q gedeeld door het geobserveerde percentage leerlingen in categorie q . Als w_q voor alle leerlingen in de normeringssteekproef bepaald is, wordt binnen elke school een steekproef met teruglegging getrokken. Bij het trekken van de steekproef wordt rekening gehouden met w_q . De trekking wordt beëindigd op het moment dat het geselecteerde leerlingaantal gelijk is aan het oorspronkelijke leerlingaantal. De steekproeftrekking wordt per school uitgevoerd, omdat het met het oog op de schoolnormering noodzakelijk is dat de scholen qua omvang en samenstelling zoveel mogelijk intact blijven (zie paragraaf 3.6). Dit is ook de reden dat in de eerste stap uitsluitend gehele scholen geselecteerd worden en geen individuele leerlingen.

Samenvattend gaat het algoritme voor het genereren van een representatieve normeringssteekproef op basis van een normeringsonderzoek (S) en Cito dataretour (D) dus als volgt te werk:

Vorbereitung data normeringsonderzoek

```
bereken  $w_{ijk}$  voor  $S$ 
indien  $w_{ijk} < 0$ 
  herhaal
    trek aselect een school  $y$  en verwijder deze uit  $S$ 
    bereken  $w_{ijk}$ 
  totdat  $w_{ijk} \geq 0$ 
retourneer  $S$ 
```

Toevoegen data uit Cito dataretour

```
bereken  $w_{ijk}$  voor  $S$ 
herhaal
  trek een school  $y$  uit  $D$  gegeven  $w_{ijk}$  en voeg deze toe aan  $S$ 
  bereken  $w_{ijk}$ 
  indien  $w_{ijk} < 0$ 
    verwijder school  $y$  uit  $S$ 
  bereken  $w_{ijk}$ 
  totdat  $N_S \geq N_T$ 
retourneer  $S$ 
```

Check leerlingvariabele sekse

```
bereken  $w_q$  voor  $S$ 
voor elke school  $y$ 
  herhaal
    trek een leerling uit  $S_y$  gegeven  $w_{y,q}$  en voeg deze toe aan  $\tilde{S}_y$ 
  totdat  $N_{\tilde{S}_y} = N_{S_y}$ 
retourneer  $\tilde{S}$ 
```

Het algoritme is toegepast bij de ontwikkeling van LVS 3.0 Begrijpend lezen. Het uitgangspunt was om de data die tijdens het *embedded field* normeringsonderzoek verzameld zijn te verdubbelen met behulp van data uit Cito dataretour. Het gewenste aantal leerlingen werd dus voor afnamemoment medio groep 4

ingesteld op $N_T = 2 \times 1673 = 3346$ en voor eind groep 4 op $N_T = 2 \times 2417 = 4834$. In tabel 4.4 is te zien in welke aantallen scholen en leerlingen het selectiealgoritme heeft geresulteerd. De conclusie is dat het voor afnamemomenten medio groep 4 en eind groep 4 tot de gewenste oplossing heeft geleid. De aantallen leerlingen die via het *embedded field* normeringsonderzoek en uit dataretour bij de normering zijn betrokken wijken weliswaar enigszins af van de nagestreefde 50:50 verhouding, maar dit is een gevolg van het exacte verloop van het algoritme gegeven de verdeling van scholen over de categorieën in de achtergrondvariabelen. Lichte afwijkingen zijn daarbij te verwachten. Dat geldt ook voor de eventuele afwijkingen in de steekproef van de populatieverdelingen voor de variabelen *regio*, *urbanisatiegraad*, *schooltype* en *geslacht*. Ook in een volledig aselechte steekproef zijn dit soort afwijkingen immers per definitie toe te schrijven aan toeval. Niettemin is in een vervolgstap de landelijke representativiteit van de normeringssteekproef ter controle onderzocht. Deze controleanalyses worden gerapporteerd in paragraaf 4.3.2.

Tabel 4.4 Aantal leerlingen per afnamemoment die meegenomen zijn in de normering

Afnamemoment	Aantal leerlingen			Aantal scholen
	Steekproef	Dataretour	Totaal normering	Normering
M4	1563	1787	3350	152
E4	2305	2554	4859	199

Nagegaan is ook of de groep scholen die geselecteerd zijn met de gegevens uit Cito dataretour een goede afspiegeling vormen van de landelijke populatie. Dat de scholen een goede afspiegeling vormen van de landelijke populatie blijkt wel uit het feit dat de gemiddelde score op de Cito Eindtoets Basisonderwijs voor deze groep niet afwijkt van het populatiegemiddelde.

4.3.2 Representativiteit

Door de werkwijze die wordt gevolgd tijdens de normering is representativiteit van de normeringssteekproeven in principe gegarandeerd. Niettemin wordt er een controle uitgevoerd op de representativiteit door de populatieverdelingen verkregen uit gegevens van DUO te vergelijken met de steekproefverdelingen. In tabel 4.5 en 4.6 worden de resultaten van de representativiteitsanalyses getoond. De steekproef is geanalyseerd in relatie tot de variabelen regio, urbanisatiegraad, schooltype en sekse.

Tabel 4.5 Representativiteitsanalyse LVS 3.0 Begrijpend lezen M4

Variabele	Categorie	Populatie (%)	Steekproef M4	
			N	%
Regio	Noord	10.1	339	10.1
	Oost	22.6	765	22.8
	West	47.4	1617	48.3
	Zuid	19.9	629	18.8
Urbanisatiegraad	Platteland	55.8	1885	56.3
	Stad	44.2	1465	43.7
Schooltype	[0, .10]	62.4	2142	63.9
	[.10, .25]	25.4	855	25.5
	[.25, .40]	6.2	183	5.5
	[.40, 1]	6.0	170	5.1
Sekse	Jongen	50.4	1554	50.5
	Meisje	49.6	1521	49.5

Tabel 4.6 Representativiteitsanalyse LVS 3.0 Begrijpend lezen E4

Variabele	Categorie	Populatie (%)	Steekproef E4	
			N	%
Regio	Noord	10.1	485	10.0
	Oost	22.6	1123	23.1
	West	47.4	2310	47.5
	Zuid	19.9	941	19.4
Urbanisatiegraad	Platteland	55.8	2751	56.6
	Stad	44.2	2108	43.4
Schooltype	[0, .10]	62.4	3119	64.2
	[.10, .25]	25.4	1225	25.2
	[.25, .40]	6.2	246	5.1
	[.40, 1]	6.0	269	5.5
Sekse	Jongen	50.4	2261	50.4
	Meisje	49.6	2229	49.6

Te zien valt dat voor de afnamemomenten M4 en E4 de steekproefverdeling weinig afwijkt van de populatieverdeling. Dit is ook statistisch getoetst. De gegevens van deze statistische toetsen worden getoond in tabel 4.7 en 4.8. De chi kwadraat-waarden zijn laag en voor elk afnamemoment in slechts één geval significant. Bij grotere steekproeven zegt significantie echter niet zoveel. Het is beter om de effectgrootte phi als uitgangspunt te nemen. We zien dat de effectgroottes ver onder de .10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). Met .052 en .055 is de effectgrootte het grootst voor de variabele schooltype voor respectievelijk afnamemoment M4 en afnamemoment E4. De conclusie luidt dat de normeringssteekproef voor de afnamemomenten M4 en E4 een zeer goede afspiegeling vormt van de populatie.

Tabel 4.7 Statistische toetsings representativiteitsanalyse LVS 3.0 Begrijpend lezen M4

Variabele	Steekproef M4				
	χ^2	df	N	p	ϕ
Regio	2.766	3	3350	.429	.029
Urbanisatiegraad	.289	1	3350	.591	.009
Schooltype	9.143	3	3350	.027	.052
Sekse	.016	1	3075	.899	.002

Tabel 4.8 Statistische toetsings representativiteitsanalyse LVS 3.0 Begrijpend lezen E4

Variabele	Steekproef E4				
	χ^2	df	N	p	ϕ
Regio	1.361	3	4859	.715	.017
Urbanisatiegraad	1.289	1	4859	.256	.016
Schooltype	14.500	3	4859	.002	.055
Sekse	.008	1	4490	.929	.001

4.3.3 Normeringsresultaten

Na de hierboven beschreven procedure doorlopen te hebben en de normeringssteekproef te hebben samengesteld kon de normering worden bepaald. Naast het gemiddelde werden de percentielen bepaald. Dat gebeurde op basis van de verdeling van scores die werden gevonden in de normeringssteekproef zoals die is samengesteld op basis van het *embedded field* normeringsonderzoek en Cito dataretour. Om de scores die leerlingen behalen te kunnen vergelijken over de tijd worden vaardigheidsscores gebruikt. Uit de ruwe scores van de leerlingen uit het *embedded field* normeringsonderzoek en Cito dataretour worden "plausible values" gegenereerd op de nieuw ontwikkelde vaardigheidsschaal. Deze "plausible values" representeren de volledige range aan vaardigheidsscores die een leerling zou kunnen hebben, gegeven de scores. De "plausible values" geven niet alleen informatie over de geschatte vaardigheid maar ook over de onzekerheid die bij die schatting hoort (Keuning et al., 2015). De normering wordt vervolgens gebaseerd op de "plausible values" van de leerlingen in de normeringssteekproef. In paragraaf 3.3 is de verdeling van "plausible values" voor de afnamemomenten M4 en E4 te zien. De "plausible values" voor deze afnamemomenten vormen een normale verdeling. Op basis van deze scoreverdeling worden de percentielen berekend die horen bij de vaardigheidsindelingen A tot en met E en I tot en met V zoals beschreven in hoofdstuk 2. Daarbij wordt uitgegaan van de empirische cumulatieve verdelingsfunctie. Volgens een transformatie worden deze percentielen op de overkoepelende schaal over leerjaren geplaatst.

Tabel 4.9 Normtabel op leerlingniveau voor LVS 3.0 Begrijpend lezen groep 4

Tijd	M	SD	K	S	P10	P20	P25	P40	P50	P60	P75	P80
medio 4	133.83	25.48	-.04	.04	101.03	112.31	116.80	127.34	133.81	140.17	150.87	155.18
einde 4	138.77	24.92	-.11	.06	107.05	117.42	121.63	132.04	138.43	144.95	155.82	160.03

Naast een normering op leerlingniveau kent Cito ook een normering op schoolniveau. Om de schoolverdeling te bepalen wordt het intercept-only multilevel model gebruikt met een gemiddelde per school en een variantie op school- en leerlingniveau. De schatting van het model verloopt via een bootstrap procedure. Dit betekent dat het multilevel model meerdere keren wordt geschat, steeds op basis van een andere selectie van scholen en leerlingen uit de normeringssteekproef. Bij elke replicatie wordt het aantal scholen dat geselecteerd gaat worden gelijkgesteld aan het aantal scholen dat in de normeringssteekproef zit. Vervolgens worden binnen een school leerlingen geselecteerd. Ook dit aantal leerlingen dat geselecteerd gaat worden, wordt gelijkgesteld aan het aantal leerlingen dat feitelijk op de betreffende school zit. De scholen en leerlingen worden geselecteerd met teruglegging. Als de selectie is afgerond, wordt het multilevel model geschat en de intraklassecorrelatie en het design effect uitgerekend. Tabel 4.10 en 4.11 laten de resultaten van de bootstrap procedure zien voor respectievelijk afnamemoment M4 en E4. De uitkomsten zijn behoorlijk stabiel. De intraklassecorrelatie (ICC) ligt boven .04 wat inhoudt dat een multilevelanalyse zinvol is (Snijders & Bosker, 1999).

Tabel 4.10 *Uitkomsten multilevelanalyse voor LVS 3.0 Begrijpend lezen afnamemoment M4*

Replicatie	Aantal		Gemiddelde	SD		ICC
	scholen	leerlingen		School	Leerling	
1	157	3374	122.420	10.074	23.962	.150
2	157	3429	121.873	9.391	23.346	.139
3	157	3358	122.460	11.693	23.425	.199
4	157	3381	121.373	9.473	22.810	.147
5	157	3176	122.260	9.910	23.432	.152
...						
15	157	3341	122.284	8.728	23.875	.118
16	157	3597	123.170	9.656	22.573	.155
17	157	3571	123.291	10.622	24.124	.162
18	157	3539	122.810	9.091	23.126	.134
19	157	3324	122.735	8.955	24.004	.122
20	157	3610	121.690	10.290	23.621	.159
Eindresultaat			122,355	10,009	23,391	0,155

Tabel 4.11 *Uitkomsten multilevelanalyse voor LVS 3.0 Begrijpend lezen afnamemoment E4*

Replicatie	Aantal		Gemiddelde	SD		ICC
	scholen	leerlingen		School	Leerling	
1	203	5032	127.064	10.467	22.980	.172
2	203	4693	126.549	10.506	23.723	.164
3	203	4796	127.046	8.476	23.327	.117
4	203	5022	126.151	9.605	23.252	.146
5	203	4554	126.409	10.232	22.998	.165
...						
15	203	4697	125.917	11.016	23.925	.175
16	203	5057	126.472	9.453	23.256	.142
17	203	4491	127.232	10.976	22.934	.186
18	203	4412	127.313	9.774	23.481	.148
19	203	4790	128.503	9.551	23.567	.141
20	203	4813	127.429	8.615	23.529	.118
Eindresultaat			126.810	9.734	23.570	.146

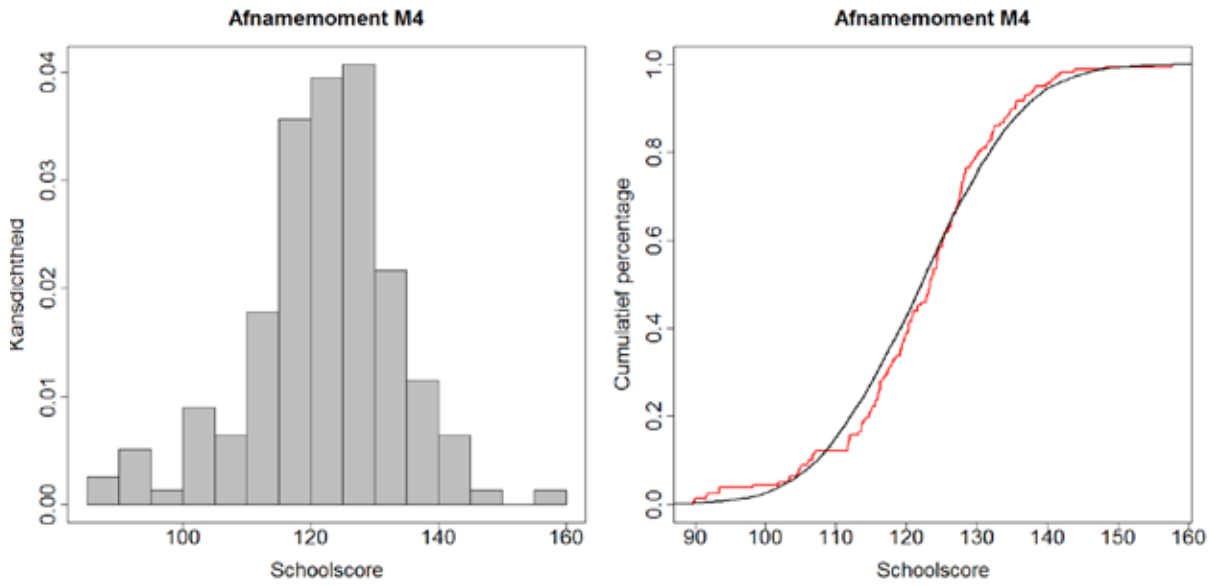
Figuur 4.7 en 4.8 laten de verdeling van schoolgemiddelden M4 en E4 zien. Het is lastig te bepalen of de schoolgemiddelden een normale verdeling volgen met een scholenaantal van 157 en 203. Op het eerste gezicht lijkt er sprake te zijn van een normale verdeling. Op basis van de resultaten van de bootstrap procedure (vergelijk de onderste regel in tabel 4.10 en 4.11) zijn de percentielen voor de vaardigheidsverdeling A tot en met E en I tot en met V berekend. Tabel 4.12 geeft de normgegevens op schoolniveau. De percentielen komen dichter bij elkaar te liggen dan in de leerlingverdeling. De afstanden zijn echter nog wel groot genoeg om scholen zinvol te classificeren in de verschillende niveaus. Het verschil tussen de vaardigheidsscores zoals weergegeven in figuur 4.7 en figuur 4.8 en die in tabel 4.12 is te verklaren doordat de scores in figuur 4.7 en figuur 4.8 worden weergegeven op de kalibratieschaal voor groep 4 terwijl de gemiddelden en standaarddeviaties in tabel 4.12 op de (getransformeerde) overkoepelende vaardigheidsschaal worden weergegeven.

Tabel 4.12 Normtabel op schoolniveau voor LVS 3.0 Begrijpend lezen groep 4

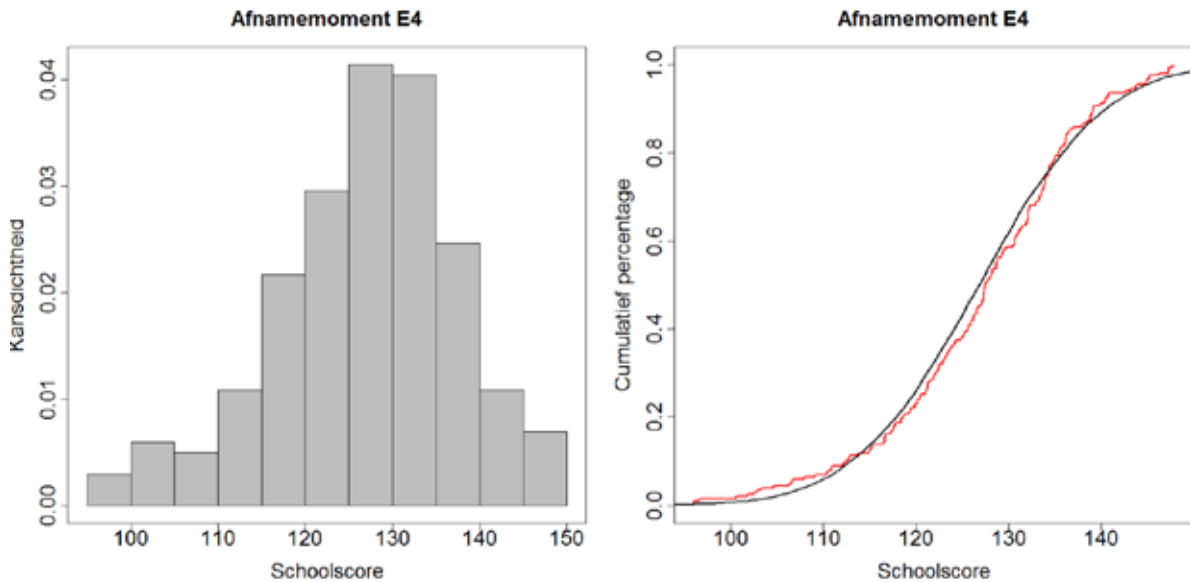
Tijd	M	SD	P10	P20	P25	P40	P50	P60	P75	P80
Medio 4	133.39	10.62	119.78	124.46	126.23	130.70	133.39	136.08	140.55	142.33
Einde 4	138.12	10.33	124.89	129.43	131.15	135.50	138.12	140.73	145.08	146.81

Noot: K = Kurtosis, S = Skewness

Figuur 4.7 Verdeling van de schoolgemiddelden voor LVS 3.0 Begrijpend lezen M4



Figuur 4.8 Verdeling van de schoolgemiddelden voor LVS 3.0 Begrijpend lezen E4



4.3.4 Geldigheid van de normen

De toetsen van het Cito Volgsysteem primair en speciaal onderwijs worden elke acht tot tien jaar vernieuwd. Niet alleen de inhoud wordt volledig vernieuwd en aangepast aan de ontwikkelingen in het

onderwijs, ook worden de normen opnieuw vastgesteld. Omdat er enige tijd verloopt tussen de dataverzameling in het normeringsonderzoek en het moment waarop een vernieuwde toets wordt uitgebracht, kan men voor de toetsen Begrijpend lezen 3.0 groep 4 een geldigheid aanhouden tot en met 2023.

Daarnaast monitort Cito periodiek de normering: jaarlijks wordt aan de hand van representatieve afnamedata nagegaan of er systematisch verschuivingen in het prestatieniveau plaatsvinden. Indien nodig wordt de normering aangepast.

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Betrouwbaarheid

Het is mogelijk om de betrouwbaarheid van de toetsen Begrijpend lezen E3M4, M4, M4E4 en E4 te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de toetsen tijdens de kalibratie OPLM-geschaald zijn. Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het feit dat de toetsen volledig bestaan uit OPLM-gekalibreerde items. Om relevante beschrijvende gegevens bij de toetsen te genereren, is gebruikgemaakt van het programma OPLAT (Verhelst, Glas en Verstralen, 1995).

In OPLAT wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ . Deze verwachte waarde wordt aangeduid met $\tau(\theta)$. Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie worden aangegeven met het symbool $Var(\tau)$. Tussen θ en $\tau(\theta)$ bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de toetsscore $\tau(\theta)$ moet behalen (dat is alleen zo als de toets oneindig lang wordt). De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor met een eenmalige toetsafname niet meer zonder fout de waarde van θ bepaald kan worden. De variantie van de geobserveerde toetsscore wordt aangegeven met $Var(t|\tau(\theta))$, en door weer gebruik te maken van de distributie van θ in de populatie kan ook de gemiddelde variantie van de geobserveerde toetsscores berekend gaan worden.

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

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

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

waarin MAcc staat voor 'Accuracy of Measurement'.

Tabel 5.1 bevat informatie over de meeteigenschappen van de vaardigheidsschaal Begrijpend lezen. In de tweede kolom staat de maximumscore, voor de toetsen is deze gelijk aan het aantal opgaven dat deel uitmaakt van de totale toetsen E3M4, M4, M4E4 en E4. De derde kolom geeft de geschatte gemiddelde score van leerlingen op de toetsen. De vierde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van de toetsen. De vijfde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de toetsen (of toetsonderdelen) is.

Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen LVS Begrijpend lezen) geeft de COTAN (COmmissie TestAangelegenheden Nederland van het Nederlands Instituut van Psychologen) aan dat een betrouwbaarheidscoëfficiënt lager dan .70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen .70 en .80 voldoende, en een betrouwbaarheidscoëfficiënt hoger dan .80 goed (COTAN Beoordelingssysteem voor de kwaliteit van tests, Evers et al., 2010, p. 33). Op grond van dit criterium is de meetnauwkeurigheid van de toetsen goed te noemen.

Tabel 5.1 Beschrijvende gegevens bij de toetsen Begrijpend lezen E3M4, M4, M4E4 en E4

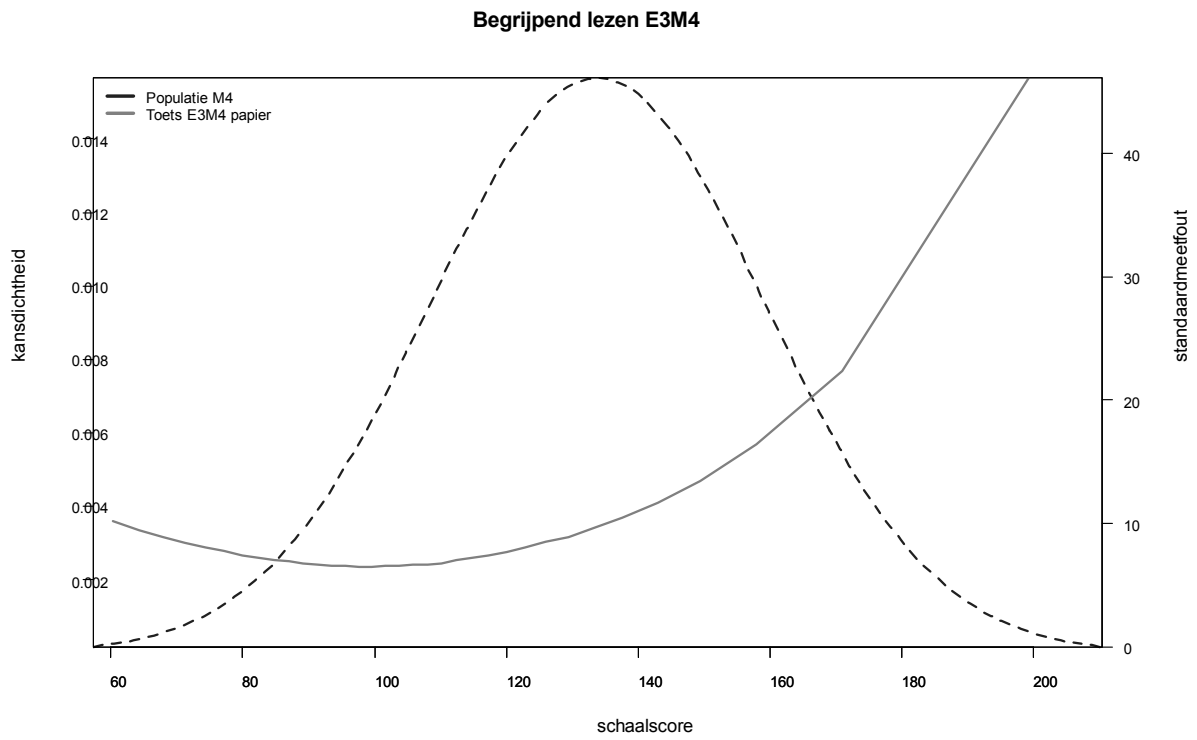
Toets	Maximum score	Gemiddelde	Standaard-meetfout	MAcc	Test-hertest (simulatie)
E3M4	40	31.7	2.285	.90	.90
M4	40	28.2	2.604	.86	.86
M4E4	40	29.4	2.537	.86	.86
E4	40	28.1	2.580	.88	.88

Er heeft geen test-hertest onderzoek plaatsgevonden. De afnamecontext van de LVS-toetsen Begrijpend lezen leent zich daar niet goed voor. Het feit dat alle items echter OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1.000.000 leerlingen. Daarbij hebben we enerzijds de vaardigheidsverdeling van alle leerlingen, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselekt uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 1.000.000 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 toetsafname en wordt niet beïnvloed door de kennis die de leerling mogelijk verworven heeft via de eerste toetsafname. Daarnaast is er sprake van invloed van een test-hertest-interval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden. De resultaten zijn weergegeven in tabel 5.1 (zie kolom 6). De uitkomst komt exact overeen met de eerder berekende coëfficiënt en leidt dan ook tot dezelfde conclusie met betrekking tot de betrouwbaarheid van de Begrijpen lezen-toets.

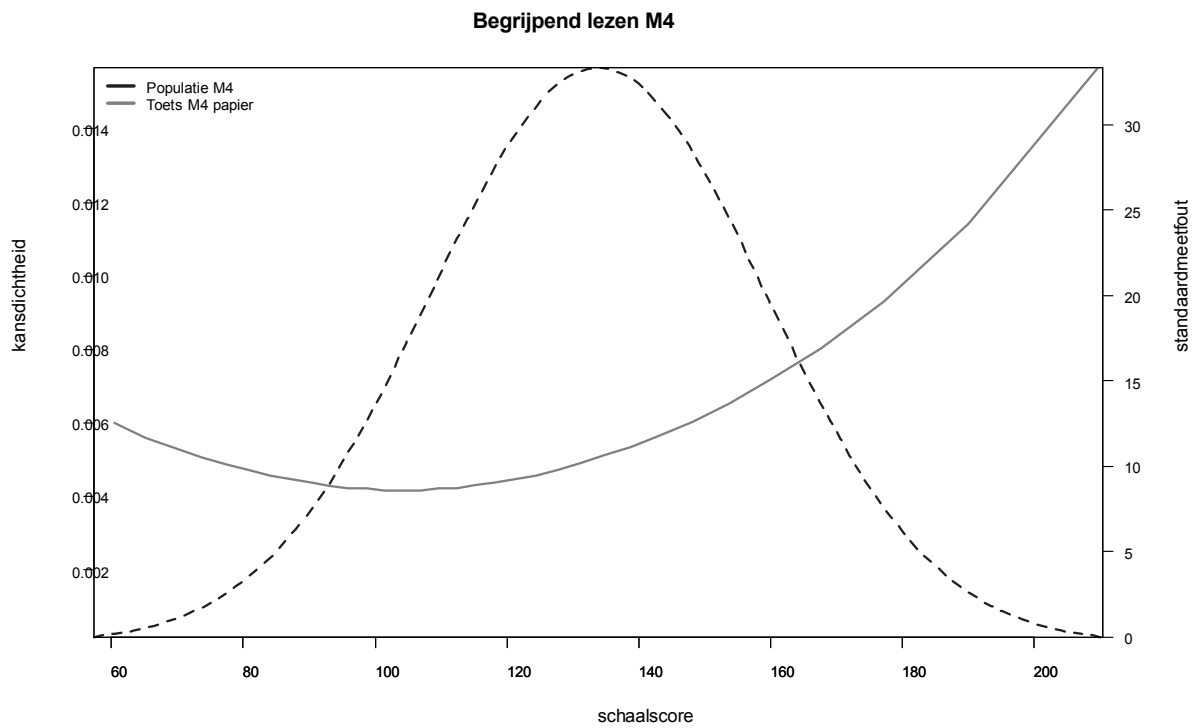
5.2 Nauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen Begrijpend lezen E3M4, M4, M4E4 en E4 en geven geen beeld van de lokale meetnauwkeurigheid van de toetsen. Figuren 5.1 tot en met 5.4 geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij deze toetsen. In deze figuren staat voor de toetsen de grootte van de meetfout op de vaardigheidsschaal afgebeeld. Ook is de kansdichtheidfunctie voor de normgroep op het afnamemoment opgenomen. Deze laat geeft weer 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 gemiddelde vaardigheidsregionen dan in de hogere vaardigheidsregionen.

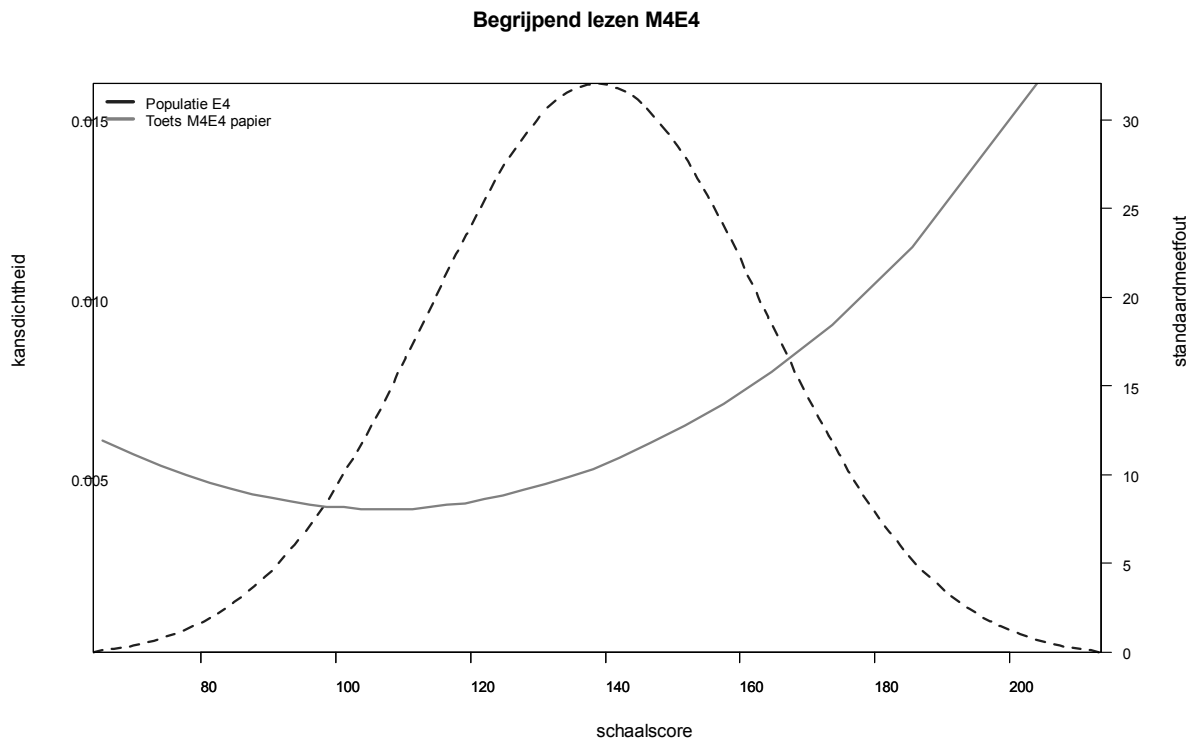
Figuur 5.1 Grootte van de meetfouten voor de toets E3M4 en de kansdichtheidsfunctie voor de M4-populatie



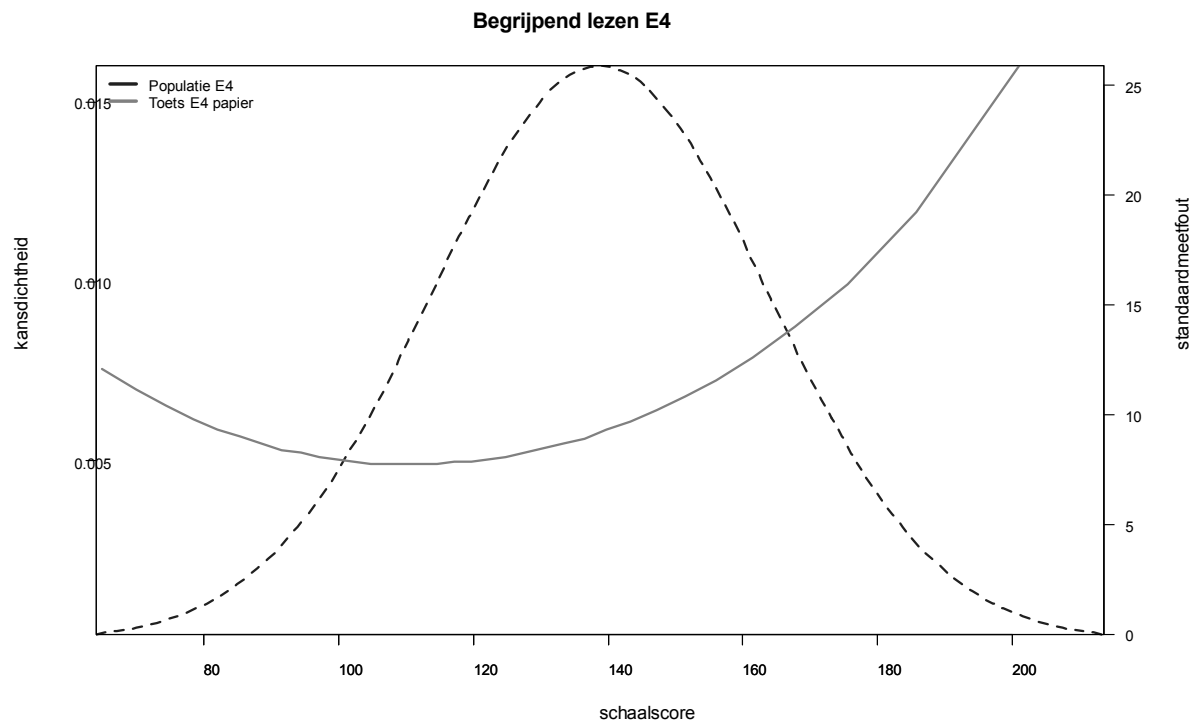
Figuur 5.2 Grootte van de meetfouten voor de toets M4 en de kansdichtheidsfunctie voor de M4-populatie



Figuur 5.3 Grootte van de meetfouten voor de toets M4E4 en de kansdichtheidsfunctie voor de E4-populatie



Figuur 5.4 Grootte van de meetfouten voor de toets E4 en de kansdichtheidsfunctie voor de E4-populatie



Voor Begrijpend lezen 3.0 groep 4

De betekenis van de (lokale) meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit betrouwbaarheidstabellen. De tabellen 5.2 tot en met 5.5 laten voor de toetsen E3M4, M4, M4E4 en E4, respectievelijk afnamemomenten medio groep 4 en einde groep 4, zien hoe vaak de werkelijke

vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat Tabel 5.2 zien dat 89,5 procent van de leerlingen die halverwege groep 4 op basis van de E3M4-toets in scoregroep V geïnclassificeerd wordt ook met hun werkelijke vaardigheidsscore in deze scoregroep geïnclassificeerd wordt. De kans dat een V-leerling terecht als

V-leerling wordt bestempeld is, met andere woorden, ongeveer 90 procent. Verder laat de linkerkant van Tabel 5.6a zien dat 10,4 procent van de leerlingen in scoregroep V een vaardigheidsscore heeft die in werkelijkheid in scoregroep IV valt. De overige getallen in Tabellen 5.2 tot en met 5.5 zijn op dezelfde wijze te interpreteren.

Tabel 5.2 Betrouwbaarheidstabel Toets E3M4 voor afnamemoment medio 4

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	89,5	10,4	0,1	0,0	0,0	E	87,1	12,8	0,1	0,0	0,0
IV	19,4	61,3	18,2	1,1	0,0	D	15,5	65,3	19,1	0,1	0,0
III	1,4	25,7	48,1	22,8	1,9	C	0,4	18,5	63,2	17,5	0,5
II	0,2	5,6	24,9	43,8	25,4	B	0	1,3	25,0	52,5	21,1
I	0,2	1,7	7,2	22,1	68,8	A	0	0,3	4,9	23,2	71,5

Tabel 5.3 Betrouwbaarheidstabel Toets M4 voor afnamemoment medio 4

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	87,0	12,5	0,5	0,0	0,0	E	83,7	15,7	0,6	0,0	0,0
IV	22,2	55,7	20,1	1,9	0,0	D	19,4	57,6	22,5	0,5	0,0
III	2,0	26	46,0	23,7	2,2	C	0,9	20,1	58,9	19,3	0,8
II	0,2	5,2	24,6	45,0	25,0	B	0,0	1,5	24,8	52,3	21,5
I	0,0	0,8	5,4	22,1	71,6	A	0,0	0,1	3,4	22,4	74,1

Tabel 5.4 Betrouwbaarheidstabel Toets M4E4 voor afnamemoment einde 4

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	87,2	12,4	0,4	0,0	0,0	E	84,1	15,3	0,6	0,0	0,0
IV	21,5	57,1	19,2	2,2	0,0	D	20,1	57,1	22,4	0,4	0,0
III	2	27	43,6	25,5	1,9	C	1	19,7	59,7	18,8	0,7
II	0,3	6,1	23,4	46,9	23,3	B	0,0	1,5	24,9	51,9	21,7
I	0,1	1,0	5,2	23	70,6	A	0,0	0,2	3,9	22,7	73,2

Tabel 5.5 Betrouwbaarheidstabel Toets E4 voor afnamemoment einde 4

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	87,4	12,3	0,3	0,0	0,0	E	84,2	15,2	0,5	0,0	0,0
IV	20,4	59,2	18,8	1,6	0,0	D	19,4	58,3	22	0,3	0,0
III	1,3	26,0	47,0	24,5	1,1	C	0,7	18,6	62,3	17,9	0,4
II	0,1	4,3	23,2	51,4	21,0	B	0,0	0,9	23,5	56	19,6
I	0,0	0,4	3,6	23	73,0	A	0,0	0,0	2,3	22	75,6

In de onderzoeksliteratuur is weinig geschreven over de beoordeling van betrouwbaarheidstabellen. Wanneer een betrouwbaarheidstabel als goed of voldoende kan worden beschouwd is onduidelijk en wat verwacht mag worden onder ideale omstandigheden is, voor zover ons bekend, niet onderzocht. Daarom worden betrouwbaarheidstabellen vaak samengevat in één of meerdere indices. Wij gebruiken de *plus/minus 1 niveau-index* en de *Marginal Classification Accuracy*. De eerste maat is bedacht door Pilliner (1969). Hij stelt als ambitieniveau dat 95 procent van de leerlingen in een scoregroep in werkelijkheid ook in die scoregroep moet scoren, **of** één scoregroep daarboven **of** één scoregroep daaronder. In de tabellen zijn dit de gearceerde cellen. Dit ambitieniveau is gebaseerd op de veronderstelling dat geen enkele toets perfect meet en dat er dus altijd sprake is van foutieve classificaties. In dat licht is de maximale accuraatheid die op het individuele niveau bereikt kan worden plus of minus één scoregroep. De tweede maat wordt op verschillende plekken in de literatuur beschreven. De maat laat zien hoe vaak de classificatie op basis van de geschatte vaardigheidsscore gemiddeld gezien overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. Bij een ideale, maar in de praktijk niet te realiseren, toetsafname lijkt de *Marginal Classification Accuracy* rond 0,75 - 0,80 uit te komen (Keuning & Béguin, in voorbereiding). In de praktijk liggen de waarden vaak tussen 0,60 en 0,70.

De samenvattende indices voor afnamemomenten medio groep 4 en einde groep 4 zijn te vinden in Tabel 5.6. Waar de betrouwbaarheidstabellen laten zien dat de meeste leerlingen op basis van hun geschatte vaardigheidsscore geplaatst worden in de niveaugroep waar ze werkelijk thuishoren, maakt Tabel 5.6 aannemelijk dat de uitkomsten duidelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969) of zelfs boven dit ambitieniveau uitstijgen. Gemiddeld gezien scoort, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 97,2 tot 99,1 procent van de leerlingen in een scoregroep ook in werkelijkheid in die scoregroep, **of** één scoregroep daarboven **of** één scoregroep daaronder. De *Marginal Classification Accuracy* loopt uiteen van 64 tot 67 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij beide afnamemomenten gemiddeld gezien in ruim 60 à 70 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. De resultaten stemmen hiermee tot grote tevredenheid: het percentage misclassificaties is erg beperkt. De laagste waarden zien we bij toets E3M4, en dan met name bij de hoogste scoregroep. Dat is conform verwachting, aangezien deze toets – die wat makkelijker is dan de toets M4 – expliciet bedoeld is voor de minst vaardige leerlingen halverwege groep 4. De (boven)gemiddeld vaardige leerlingen zullen deze toets in de praktijk ook niet maken.

Op basis van bovenstaande gegevens concluderen we dat op basis van de toetsen Begrijpend lezen 3.0 groep 4 de leerlingen op een betrouwbare manier ingedeeld kunnen worden in normgroepen. Deze indeling voldoet uitstekend gegeven het doel van de toets. Uiteraard dienen de gebruikers rekening te houden met het gegeven dat er altijd sprake zal zijn van misclassificatie; veelal van maximaal één niveau verschil.

Tabel 5.6 *Samenvattende indices toetsen E3M4, M4, M4E4 en E4 op afnamemomenten groep 4*

	Toets E3M4, afnamemoment M4		Toets M4, afnamemoment M4		Toets M4E4, afnamemoment E4		Toets E4, afnamemoment E4	
	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	66,3	67,0	64,1	63,9	64,6	64,1	66,7	66,2
Accuracy plus/minus 1 niveau	97,4	98,8	97,3	98,6	97,2	98,5	98,1	99,1

6 Validiteit

De begripsvaliditeit van een toets gaat over 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. Bij leervorderingstoetsen, zoals deze toetsen Begrijpend lezen voor groep 4, speelt de inhoudsvaliditeit een relatief belangrijke rol, meer wellicht dan bij psychologische testen in het algemeen. In paragraaf 6.1 wordt beschreven waarop de inhoudsvaliditeit van de toets gebaseerd is; daarbij grijpen we terug op de inhoudsverantwoording zoals deze in hoofdstuk 3 is beschreven. De paragrafen 6.2 tot en met 6.6 zijn gewijd aan een aantal aspecten van begripsvaliditeit. In paragraaf 6.2 wordt het unidimensionele karakter van de toets aangegeven en worden gegevens over de structuur van de toets gepresenteerd. In paragraaf 6.3 wordt de kwaliteit van het itemmateriaal behandeld. Paragraaf 6.4 gaat over onderzoek naar itembias. Paragraaf 6.5 behandelt het soortgenootonderzoek dat in het kader van de ontwikkeling van deze toets is uitgevoerd. Dit onderzoek levert data op voor de convergente en divergente validiteit. Als laatste komen in paragraaf 6.6 verschillen tussen relevante groepen aan bod.

6.1 Inhoudsvaliditeit

De samenstelling van de toets is bepaald door inhoudelijke criteria en psychometrische criteria. Voor de inhoudsvaliditeit zijn de inhoudelijke criteria relevant. Inhoudelijk zijn richtinggevend geweest het referentiekader Nederlandse taal (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a), de kerndoelen Nederlandse taal (Ministerie van OCW, 2006), de tussendoelen gevorderde geletterdheid voor de middenbouw (Aarnoutse & Verhoeven, 2003), de Leerstoflijnen lezen (Oosterloo & Paus, 2010) en recente wetenschappelijke publicaties over begrijpend lezen.

Deze bronnen vormen de basis voor de domeinbeschrijving van de toetsen Begrijpend lezen, die aan bod komt in hoofdstuk 2. In hoofdstuk 3 is de domeinbeschrijving uitgewerkt in een beschrijving en verantwoording van tekstsoorten, opgavenvormen en vaardigheden binnen de toetsen Begrijpend lezen. De constructie van de opgaven is eveneens afgeleid van deze domeinindeling en ook de definitieve selectie van opgaven in de toetsen is gebaseerd op een gewenste verdeling van verschillende typen opgaven binnen en over de verschillende domeinen. Beoogd is de toetsen onafhankelijk samen te stellen van de verschillende onderwijsmethoden. Bij de constructie van de opgaven zijn leerkrachten uit het onderwijs betrokken zodat de opgaven voor wat betreft leesbegrip aansluiten bij leerlingen van groep 4. Uit de manier waarop de inhoudelijke criteria zijn toegepast, kunnen we dan ook concluderen dat de inhoudsvaliditeit van de toetsen Begrijpend lezen voor E3M4, M4, M4E4 en E4 goed te noemen is.

6.2 Unidimensionaliteit, respectievelijk structuur

Unidimensionaliteit van de te meten vaardigheid is het uitgangspunt van het meetmodel van de toetsen Begrijpend lezen. Naast de inhoudelijk relevante indeling van de opgaven is een unidimensionele schaal gerealiseerd. Dit betekent dat met elke willekeurige subset van items dezelfde onderliggende vaardigheid kan worden vastgesteld.

Zoals in hoofdstuk 4 al aangegeven zijn bij de kalibratie voor alle toetsopgaven S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij is duidelijk geworden dat de verdeling van overschrijdingskansen bij deze statistische toetsen gelijkmatig is over het gehele interval waarin de overschrijdingskansen kunnen liggen (i.e. tussen 0 en 1). Dit resultaat geeft een bevestiging van het eerder geschetste beeld, dat er met uitzondering van enkele opgaven, sprake is van niet-significante S-toetsen. Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensioneel construct representeren (zie tabel 4.1).

Ook in hoofdstuk 4 zijn als maat voor de modelfit de R1c-waarden gepresenteerd. Omdat deze eveneens ondersteuning bieden voor de validiteit refereren we daar nogmaals aan. R1c is een statistiek die zicht geeft op de modelpassing van de toets als geheel. Voor een acceptabele modelfit geldt als vuistregel dat R1c bij voorkeur niet significant zou moeten zijn en niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden (df). In tabel 4.2 zijn deze waarden te vinden.

De modelpassing van de toetsen voldoet aan de voorwaarde dat R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootte is significant voor de toetsen, maar hier moet bij steekproeven van deze omvang niet teveel waarde aan worden gehecht.

Ten slotte kan de nauwkeurigheid van de itemparameterschattingen (aan de hand van de constante 'c', zie hierover het COTAN Beoordelingssysteem; Evers, Lucassen, Meijer & Sijtsma, 2010, p 40) als uitstekend worden beoordeeld. In hoofdstuk 4 is deze informatie al weergegeven maar omdat deze ook relevant is voor de validiteit wordt deze hier nog eens aangehaald. In tabel 4.3 zijn het gemiddelde en de range van deze waarden voor de toetsitems weergegeven. De gemiddelde waarde van de constante voor de vier toetsen (variërend van .063 tot .095) is te interpreteren als uitstekend. Voor geen enkele opgave is c bovendien groter dan .20. De conclusie mag luiden dat we ook op basis van deze analyse de kalibratie geslaagd kunnen noemen.

6.3 Itemkwaliteit

In tabel 6.1 zijn de ranges en de gemiddelden weergegeven voor de p-waarden en de R_{it} -waarden van de items van de toetsen E3M4 tot en met E4. Voor de toetsen is te zien dat de p-waarden liggen tussen de .44 en .90. De p-waarden van de items liggen dus tussen de beoogde .40 en .90. Er is gezorgd voor een goede spreiding van moeilijkheid over de items. De gemiddelde moeilijkheid van de M4-, en E4-toetsen is .71 en .70, waar het streven was tussen .60 en .75. Daarmee zijn de toetsen niet te moeilijk en wordt voorkomen dat de leerling gefrustreerd raakt tijdens de toetsafname. De toetsen E3M4 en M4E4 zijn zoals beoogd iets makkelijker. De R_{it} -waarde ligt voor vier items uit de toets M4 en vijf items uit de toets M4E4 onder de .30, maar is voor alle items groter dan .20 (in het COTAN Beoordelingssysteem de ondergrens voor de beoordeling 'voldoende'). Voor de toetsen E3M4 en E4 zijn de R_{it} -waarden van alle items .30 of hoger. Door de Cotan wordt een R_{it} -waarde groter dan .30 gekwalificeerd als goed. Met gemiddelden van .39 tot .45 is de itemkwaliteit van de toetsen goed te noemen. Bijlage 3 bevat een volledig overzicht van de p-waarden en de R_{it} -waarden van de items uit de toets.

Tabel 6.1 Range en gemiddelde van p- en R_{it} -waarden toetsmomenten E3M4, M4, M4E4 en E4

	P-waarden		R_{it} -waarden		N items
	Range	Gemiddelde	Range	Gemiddeld	
E3M4	.63-.90	.79	.30-.59	.45	40
M4	.44-.90	.71	.26-.54	.39	40
M4E4	.50-.88	.74	.24-.53	.40	40
E4	.50-.85	.70	.32-.52	.42	40

In tabel 6.2 wordt de verdelingskarakteristiek gegeven van de ruwe scores op het afnamemoment dat hoort bij de toetsen E3M4, M4, M4E4 en E4. Het gemiddelde komt uiteraard overeen met wat men bij een gegeven aantal items mag verwachten bij de gekozen (gemiddelde) moeilijkheidsgraad. Omdat deze gemiddelde moeilijkheidsgraad rond .70 ligt voor de toetsen M4 en E4 en iets hoger is voor de toetsen E3M4 en M4E4, is de verdeling linksscheef (vergelijk de negatieve waarden in de kolom 'scheefheid'). De verdeling is ééntoppig.

Tabel 6.2 Verdelingskenmerken van de toets Begrijpend lezen E3M4, M4, M4E4 en E4

Meetmoment	Aantal opgaven	Gemiddelde	SD	Scheefheid	Kurtosis
E3M4	40	31.7	7.13	-1.252	1.235
M4	40	28.2	6.91	-.745	.101
M4E4	40	29.4	6.79	-.889	.405
E4	40	28.1	7.45	-.744	.025

6.4 Itembias

Er is onderzoek uitgevoerd naar differentieel itemfunctioneren (*Differential Item Functioning*, DIF) met betrekking tot sekse. Voor alle toetsopgaven zijn geobserveerde en verwachte scores voor zowel jongens als meisjes in verschillende scoregroepen berekend. Vervolgens is hier een S-statistiek voor berekend, analoog aan hoe dit gebeurt tijdens de kalibratie (zie hoofdstuk 4).

Het onderzoek naar DIF over sekse per item liet bij geen enkel item van de toetsen E3M4, M4, M4E4 en E4 differentieel functioneren zien (bij $\alpha = .01$). Voor de toetsen Begrijpend lezen van leerjaar 4 is er dus geen sprake van itembias met betrekking tot sekse.

6.5 Soortgenotenonderzoek

In het kader van het soortgenotenonderzoek is gekeken naar de convergente validiteit door de samenhang te onderzoeken met de voorgaande versies van de LVS-toetsen Begrijpend lezen (2.0) van groep 4 en eveneens met de LVS 3.0-toets Begrijpend lezen van groep 3. De divergente validiteit is onderzocht door de samenhangen met andere taalonderdelen en rekenen-wiskunde te onderzoeken.

Convergente validiteit: Correlatie scores LVS 2.0 met LVS 3.0 Begrijpend lezen

De leerlingen van de normeringssteekproeven hebben zowel opgaven van de toetsen Begrijpend lezen van LVS 2.0 als opgaven van de toetsen Begrijpend lezen LVS 3.0 gemaakt en daardoor kunnen correlaties worden bepaald tussen de vaardigheidsscores op basis van de LVS 2.0 opgaven en die op basis van de LVS 3.0 opgaven. De correlatie tussen de gewogen score gebaseerd op items uit LVS 2.0 en de gewogen score gebaseerd op items uit LVS 3.0 is hoog voor zowel M4 ($r=.86$; $N=1673$) als E4 ($r=.0,96$; $N=2421$). Aangezien de begripsvaliditeit van de toetsen Begrijpend lezen LVS 2.0 door de Cotan als 'goed' is beoordeeld, vormen deze hoge correlaties een belangrijke schakel in de bewijsvoering omtrent de validiteit van de LVS 3.0 toetsen.

Convergente validiteit: Correlatie scores LVS 3.0 E3, M4 en E4 Begrijpend lezen

Een van de functies van de LVS-toetsen is het beschrijven en volgen van de vaardigheidsgroei in begrijpend lezen van leerlingen over de jaren heen. Tegelijkertijd met LVS 3.0 Begrijpend lezen groep 4 is ook de toets LVS 3.0 Begrijpend lezen groep 3 uitgegeven. Om een uitspraak te kunnen doen over de correlaties tussen leerlingen op verschillende afnamemomenten is een gedeelte van de leerlingen uit de normeringssteekproef E3 gevolgd. Deze leerlingen hebben in schooljaar 2013-2014 meegedaan aan een onderzoek waarin de beoogde toetsen voor M4 en E4 zijn voorgelegd. Daarnaast zijn extra scholen benaderd voor het afnamemoment M4, die gevolgd zijn naar afnamemoment E4. In tabel 6.3 worden de correlaties tussen de gewogen scores op de verschillende afnamemomenten weergegeven.

Tabel 6.3 Correlaties tussen de scores op de LVS 3.0- toetsen Begrijpend lezen E3, M4 en E4

	E3	M4	E4
E3	---	915	762
M4	.82	---	1398
E4	.75	.90	---

Alle correlaties zijn met waarden van .75 of hoger hoog te noemen. De correlatie tussen de score op de M4-toets en de score op de E4-toets is het hoogst (.90), terwijl de correlatie tussen de scores op de E3- en de E4-toets het laagst is. Dit ligt in de lijn der verwachtingen aangezien de leerlingen gedurende leerjaar 3 en 4 zich niet alleen het begrijpen van teksten eigen maken, maar ook het technisch/nauwkeurig lezen nog een grote rol speelt. Naarmate leerlingen ouder worden, is dit laatste minder het geval. Tevens zit er een langere afnameperiode tussen de afnamemomenten E3 en E4, maar ook tussen E3 en M4 dan tussen M4 en E4. Het interval tussen de afnamemomenten lijkt dus van invloed op de hoogte van de correlatie.

Divergente validiteit: Correlatie scores LVS 3.0 met diverse toetsen leervorderingen

Aan de scholen die hebben deelgenomen aan het normeringsonderzoek E4 en die gebruikmaken van het computerprogramma LOVS is gevraagd of ze bereid waren gegevens van andere LVS-toetsen uit het Cito Volgsysteem aan te leveren. Dit is mogelijk via een functie van het computerprogramma die bij de scholen bekend staat onder de noemer 'dataretour'. Met de dataretourfunctie zijn van de leerlingen op deze manier ook scores op andere leervorderingstoetsen van Cito beschikbaar.

Op basis van algemene cognitieve verschillen (intelligentie) is er altijd sprake van een zekere samenhang tussen toetsscores op verschillende vakgebieden. Hoe sterk deze samenhang is voor een specifieke combinatie van toetsscores, hangt af van de betreffende vakgebieden. We verwachten dat de vaardigheid in begrijpend lezen in groep 4 redelijk sterk samenhangt met de woordenschat. In beide toetsen ligt sterk de nadruk op de semantiek en in het tekstbegrip speelt woordkennis een belangrijke rol. In vergelijking met groep 3 zal de correlatie met technisch lezen lager zijn. In groep 3 wordt het 'met begrip' lezen nog vrij sterk bepaald door de nauwkeurigheid van het (technisch) lezen, eind groep 4 is dat veel minder het geval. De goede (technische) lezers onderscheiden zich dan door hun leessnelheid. Bij de toetsen Begrijpend lezen speelt het tempo echter geen rol: de leerlingen mogen zo lang over de toetsen doen als nodig is. Ook voor spellingvaardigheid verwachten we een matige samenhang, omdat deze vaardigheid net als de vaardigheid in technisch lezen niet-semantisch van karakter is. Voor rekenen-wiskunde, ten slotte, verwachten we een matige samenhang die vergelijkbaar is met de samenhang tussen deze vakgebieden in groep 3. Aan de ene kant betreft het een ander domein, namelijk rekenen-wiskunde in plaats van taal, aan de andere kant speelt bij beide vaardigheden het analytisch vermogen een rol.

In tabel 6.4 is de correlatie tussen de scores op de toets Begrijpend lezen E4 en de toetsen Technisch lezen (Leestempo, DMT), Spelling, Rekenen-Wiskunde en Woordenschat op ditzelfde afnamemoment weergegeven. Voor de toets Begrijpend luisteren zijn ook leerlinggegevens opgevraagd, maar was de respons te laag (N=9) om correlaties te kunnen berekenen.

Tabel 6.4 Correlaties tussen Begrijpend lezen E4 en verschillende andere LVS-onderdelen

	Begrijpend lezen*	Aantal leerlingen
Cito Technisch lezen – Leestempo E4	.49	111
Cito Technisch Lezen – DMT E4	.55	287
Cito Spelling E4	.60	274
Cito Rekenen-Wiskunde E4	.55	286
Cito Woordenschat E4	.74	272

* Deze correlaties zijn gecorrigeerd voor attenuatie

Uit tabel 6.4 blijkt dat de samenhang tussen enerzijds begrijpend lezen en anderzijds woordenschat inderdaad redelijk sterk is. Het betreft de hoogste correlatie in deze tabel.

De correlatie met technisch lezen is lager in vergelijking met groep 3. In de Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 3 (Jolink, Tomesen, Hilde, Weekers en Engelen, 2015) staan de correlaties vermeld tussen Begrijpend lezen E3 en andere toetsen E3. De correlatie op het E3-moment tussen begrijpend lezen en verschillende toetsen voor technisch lezen lag tussen .62 en .71. De correlaties met spelling en rekenen-wiskunde aan het einde van groep 4 zijn in overeenstemming met onze verwachtingen en ook vergelijkbaar met de correlaties aan het einde van groep 3. Die waren toen .61 respectievelijk .56.

Samenvattend kan dus gesteld worden dat de correlaties van de LVS 3.0 – toetsen groep 4 Begrijpend lezen conform verwachting zijn. De correlatie met de vaardigheidsscores op basis van LVS 2.0 Begrijpend lezen (een soortgenoot, waarvan de begripsvaliditeit positief is beoordeeld), is hoog voor zowel afnamemoment M4 als voor afnamemoment E4. Ook de correlaties tussen toetsscores voor Begrijpend lezen 3.0 op verschillende afnamemomenten (E3, M4 en E4) zijn hoog. Voor afnamemoment E4 zien we dat de correlatie met LVS 2.0 duidelijk hoger is dan de correlatie met andere leervorderingentoetsen, die op hetzelfde moment zijn afgenomen. Deze gegevens vormen een ondersteuning voor de validiteit van de toetsen. De data geven aan dat er gemeten wordt wat men beoogt te meten, namelijk begrijpend lezen.

6.6 Verschillen tussen relevante subgroepen

Bij de normeringsonderzoeken zijn geboortedatum, geslacht en leerlinggewicht van de leerlingen opgevraagd. Voor deze drie variabelen zullen de verschillen tussen subgroepen worden besproken. Op basis van de geboortedatum is de leeftijd van de leerlingen bepaald. Leerlingen zijn vervolgens ingedeeld in leeftijdsgroepen. In tabel 6.5 wordt de gemiddelde score van de verschillende leeftijdsgroepen weergegeven.

Tabel 6.5 Gemiddelde score per leeftijdsgroep voor de afnamemomenten M4 en E4

M4

Leeftijdsgroep	Aantal	M	SD
6.5	2	165.0	19.3
7-8	1329	135.2	25.7
8.5 en ouder	170	121.9	23.0

E4

Leeftijdsgroep	Aantal	M	SD
6.5-7	7	158.2	11.3
7.5-8.5	1894	140.5	26.4
9 en ouder	263	125.6	22.7

Het patroon in de tabel is naar verwachting en komt overeen met eerder gevonden verschillen tussen reguliere en vertraagde, respectievelijk versnelde leerlingen (zie bijvoorbeeld Kuhlemeier, Jolink, Krämer, Hemker, Jongen, van Berkel & Bechger, 2014). De jongste leeftijdsgroep scoort hoog in vergelijking met de andere leeftijdsgroepen (effectgrootte = 1.17 voor M4 en effectgrootte = .73 voor E4 ten opzichte van het algemeen gemiddelde). Deze leerlingen zijn de versnelde leerlingen die op grond van hun cognitieve capaciteiten en/of leerprestaties een groep hebben overgeslagen. Let wel dat dit slechts om twee leerlingen op afnamemoment M4 en zeven leerlingen op afnamemoment E4 gaat. Aan de andere kant scoort de oudste groep leerlingen het laagst (effectgrootte -.48 voor M4 en effectgrootte -.51 voor E4). Ook hier is dat

naar verwachting aangezien deze leerlingen in veel gevallen op grond van hun leerprestaties een jaar gedoubleerd hebben. De leerlingen die zitten in de jaargroep, die op basis van de leeftijd wordt verwacht, behalen een score die valt tussen de scores van de jongere en oudere leerlingen.

Ook voor de variabele geslacht is de gemiddelde score van jongens en meisjes bepaald, zie tabel 6.6. De meisjes behalen een iets hogere score dan de jongens. Er is sprake van een klein effect (-.28 voor afnamemoment M4 en -.24 voor afnamemoment E4). Dit is naar verwachting. Het is een bekend verschijnsel dat meisjes beter scoren op toetsen begrijpend lezen dan jongens (zie bijv. Kuhlemeier et al., 2014).

Tabel 6.6 Gemiddelde score jongens en meisjes voor de afnamemomenten M4 en E4

M4

Geslacht	Aantal	M	SD
jongen	689	129.7	25.6
meisje	679	136.9	25.8

E4

Geslacht	Aantal	M	SD
jongen	1015	135.2	26.0
meisje	1006	141.6	26.6

Tenslotte is gekeken naar de variabele leerlinggewicht. De gemiddelde score van 0.00-, 0.30- en 1.20-leerlingen is bepaald en wordt weergegeven in tabel 6.7. De leerlingen met een gewicht 0.00 behalen de hoogste scores (effectgrootte .13 voor afnamemoment M4 en effectgrootte .10 voor afnamemoment E4 ten opzichte van het algemene gemiddelde). Aan de andere kant scoren de leerlingen met het hoogste gewicht (1.20) het laagst (effectgrootte -1.06 voor M4 en effectgrootte -.83 voor E4). De leerlingen met het gewicht 0.30 vallen er tussenin. Ook dit is volgens verwachting. De leerlingen met ouders met een lagere vooropleiding scoren lager op de toetsen Begrijpend lezen dan leerlingen waarvan de ouders een hogere opleiding hebben genoten (zie bijv. Kuhlemeier et al., 2014).

Tabel 6.7 Gemiddelde score per gewicht voor de afnamemomenten M4 en E4

M4

Gewicht	Aantal	M	SD
0.00	798	137.8	25.0
0.30	70	123.8	22.1
1.20	46	106.8	22.8

E4

Geslacht	Aantal	M	SD
0.00	1443	141.7	26.6
0.30	123	126.1	24.6
1.20	99	117.1	22.8

7 Samenvatting

In dit samenvattende hoofdstuk geven we kort weer wat in de voorafgaande hoofdstukken is besproken. De toetsen Begrijpend lezen 3.0 voor groep 4 vormen een hulpmiddel om vast te stellen in hoeverre leerlingen geschreven teksten begrijpen. De toetsen kunnen, in samenhang met de toetsen Begrijpend lezen 3.0 voor groep 3 en die voor de hogere leerjaren, worden gebruikt om de leesvaardigheid van leerlingen in het primair en speciaal onderwijs in kaart te brengen en om hun ontwikkeling te volgen.

We beschreven in hoofdstuk 2 dat de inhoud van de toets aansluit bij het referentiekader Nederlandse taal, de kerndoelen Nederlandse taal, de tussendoelen gevorderde geletterdheid en de Leerstoflijnen lezen. In de domeinbeschrijving legden we uit welke aspecten en deelvaardigheden een rol spelen bij begrijpend lezen en beschreven we de ontwikkeling van de vaardigheid. Daarnaast beschreven we de opgavenbanken die gebruikt worden voor de toetsen van het Cito Volgstelsel voor primair en speciaal onderwijs en lichtten we toe dat de vaardigheid begrijpend lezen kan worden opgevat als een unidimensioneel continuüm. Verder werd in hoofdstuk 2 het gehanteerde meetmodel (OPLM) beschreven, dat gebaseerd is op de itemresponstheorie.

Nadat we in hoofdstuk 2 de uitgangspunten bij de toetsconstructie hebben beschreven, is in hoofdstuk 3 de inhoud van de toets uitgewerkt. Daarbij zijn de doelen voor de toetsen van groep 4 uitvoerig beschreven en is er een vergelijking gemaakt tussen de gewenste en gerealiseerde verdeling van toetsitems. Ook is in dit hoofdstuk verslag gedaan van de itemconstructie, de opzet van de proeftoetsing en het normeringsonderzoek en de samenstelling van de definitieve toets. Ten slotte bevat hoofdstuk 3 een beknopte statistische beschrijving van de toets.

In hoofdstuk 4 rapporteerden we over de kalibratie en normering. We beschreven de opzet, de gevolgde stappen bij de kalibratie en de toetsing van het gehanteerde IRT-model. Verschillen in gedrag tussen de leerlingen zijn te verklaren door een unidimensioneel concept. Uit de resultaten van de S-toetsen op het niveau van de individuele toetsitems, de analyses in termen van R1c en de zogenoemde constante 'c' trokken we de conclusie dat de kalibratie geslaagd is.

In paragraaf 4.3.2 werd aangetoond dat de normeringssteekproef op basis van de variabelen regio, urbanisatiegraad, schooltype en sekse een zeer goede afspiegeling vormt van de populatie. De data in deze steekproeven vormen een combinatie van uitkomsten van toetsafnames in de vorm van *embedded field* onderzoek en dataretour. We betoogden dat de gekozen aanpak de best mogelijke garantie vormt voor een adequate initiële normering. In de laatste paragraaf van hoofdstuk 4 presenteerden we de normeringsresultaten en gaven we aan met welke vaardigheidsscores de grenzen van de niveau-indelingen samenvallen.

In hoofdstuk 5 staat de betrouwbaarheid van de toets centraal. De betrouwbaarheidscoëfficiënten van de vier toetsen zijn hoger dan .85 en daarmee is de betrouwbaarheid goed te noemen. Verder zijn in dit hoofdstuk betrouwbaarheidstabellen opgenomen die de betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden, laten zien. Daarnaast gaven we inzicht in de lokale betrouwbaarheid: de meetfout blijkt het kleinst te zijn in de lagere en gemiddelde vaardigheidsregionen.

In het laatste hoofdstuk, hoofdstuk 6, stelden we de inhoudsvaliditeit en de begripsvaliditeit van de toets aan de orde. De inhoudsvaliditeit werd aangetoond door te verwijzen naar de verschillende bronnen die in het Nederlands onderwijs richtinggevend zijn voor de inhoud van het domein begrijpend lezen, naar de inhoudelijke verantwoording van de opgenomen items en naar de procedures waarmee de toets is geconstrueerd. De begripsvaliditeit werd besproken door te verwijzen naar het unidimensionele karakter van de toets, zoals dat in hoofdstuk 4 is aangetoond. Uit de resultaten van de kalibratieanalyses is al af te leiden dat de kwaliteit van de items hoog is. Dit wordt bevestigd door de 'klassieke' itemparameters: zowel de p-waarden als R_{it} -waarden zijn zeer goed te noemen. DIF-onderzoek toont daarnaast aan dat er bij geen enkel item in de toets sprake is van itembias met betrekking tot sekse. De convergente en divergente

validiteit werd onderzocht. De correlaties met soortgelijke toetsen Begrijpend lezen zijn hoog, terwijl de correlaties met andere leervorderingentoetsen lager zijn. In de laatste paragraaf worden de verschillen tussen de relevante subgroepen naar leeftijd, sekse en leerlinggewicht gegeven. Ook hier zijn de resultaten in overeenstemming met de verwachtingen.

8 Literatuur

Aarnoutse, C. & L. Verhoeven (2003). *Tussendoelen gevorderde geletterdheid. Leerlijnen voor groep 4 tot en met 8*. Nijmegen: Expertisecentrum Nederlands.

Alexander, P.A., & Jetton, T.L. (2000). Learning from Text: A Multidimensional and Developmental Perspective. In: P.B. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr, (Eds.), *Handbook of Reading Research, Volume 3*. pp. 285-310. Mahwah, NJ: Lawrence Erlbaum.

Boerma, I., Manders, A. & Markhorst, M. (2009). Begrijpend lezen nieuwe stijl. In: Tjalling Brouwer (red.). *Taalspecialisten aan het werk. Een bundel artikelen geschreven in het kader van de opleiding voor taalspecialisten*, pp. 55-79. Enschede: Stichting Leerplanontwikkeling (SLO).

Boxtel, H. van & B.T. Hemker (2009). *Wetenschappelijke verantwoording van de Intelligentietest Eindtoets Basisonderwijs*. Arnhem: Cito.

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O. & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001 (2nd ed.)*. Chestnut Hill, MA: Boston College Cito (2014). *Cito Volgsysteem primair en speciaal onderwijs. Begrijpend lezen 3.0 Groep 4*. Arnhem: Cito.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.

Diepen, M. van (2007). *Variation in reading literacy; a cross-national approach* (proefschrift). Radboud Universiteit Nijmegen.

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. 309-348. Arnhem: Cito.

Elsäcker, W. (2002). *Development of Reading Comprehension: The Engagement Perspective* (dissertatie). Nijmegen: KUN.

Evers, A., Lucassen, W., Meijer, R. & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests*. Amsterdam: NIP/COTAN.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008a). *Over de drempels met taal en rekenen. Hoofdrapport van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen*. Enschede: SLO.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008b). *Over de drempels met taal. De niveaus voor de taalvaardigheid*. Enschede: SLO.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009a). *Referentiekader taal en rekenen. De referentieniveaus*. Enschede: SLO.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009b). *Een nadere beschouwing. Over de drempels met taal en rekenen*. Enschede: SLO.

Fisher, F., Frey, H. & Lapp, D. (2009). *In a reading state of mind. Brain research, teacher modeling and comprehension instruction*. Newark: International reading association.

- Förrer, M., & Van de Mortel, K. (2010). *Lezen ... denken ... begrijpen! Handboek begrijpend lezen in het basisonderwijs*. CPS onderwijsontwikkeling en advies, Amersfoort.
- 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.
- Goldman, S.R. & Rakestraw, J.A. jr. (2000). Structural Aspects of Constructing Meaning From Text. In P.B. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr, (Eds.), *Handbook of Reading Research, Volume 3*. pp. 311-335. Mahwah, NJ: Lawrence Erlbaum.
- Hemker, B.T., J. Kordes & J.J. van Weerden (2011). *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2010 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau*. Arnhem: Cito.
- Jolink, A., Tomesen, M., Hilte, M., Weekers, A. & Engelen, R. (2015). *Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 3*. Arnhem: Cito.
- Keuning, J. (2011). *Normeren op schoolniveau met Cito dataretour*. Arnhem: Cito.
- Keuning, J., Boxtel, H. van, Lansink, N., Visser, J., Weekers, A. & Engelen, R. (2015). *Actualiteit en kwaliteit van normen. Een werkwijze voor het normeren van een leerlingvolgsysteem*. Arnhem: Cito.
- Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. In R. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading, 5th ed.*, pp. 1270-1328. Newark, DE: International Reading Association.
- Kuhlemeier, H., Jolink, A., Krämer, I., Hemker, B., Jongen, I., Van Berkel, S. & Bechger, T. (2014). *Balans van de leesvaardigheid in het basis- en speciaal basisonderwijs 2. (PPON-reeks nummer 54)*. Arnhem: Cito.
- Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- McNamara, D.S. & Kendeou, P. (2011). Translating advances in reading comprehension research to educational practice. *International Electronic Journal of Elementary Education*, 4 [1], pp. 33-46.
- Ministerie van Onderwijs, Cultuur en Wetenschappen (2006). *Kerndoelenboekje*. www.minocw.nl
- Mullis, I.V.S, Kennedy, A.M., Martin, M.O. & Sainsbury, M. (2006). PIRLS 2006, *Assessment framework and specifications (2nd edition) Progress in International Reading Literacy Study*. Chestnut Hill, MA. Boston College.
- Oosterloo, A. en Paus, H. (2010). *Leerstoflijnen lezen beschreven. Uitwerking van het referentiekader Nederlandse taal voor het lesonderwijs op de basisschool*. SLO (nationaal expertisecentrum leerplanontwikkeling), Enschede.
- Pilliner, A. (1969). *Estimation of number of grades to be awarded in an examination by consideration of its reliability coefficient*. Edinburgh: The Godfrey Thomson Unit for Educational Research.
- Pressley, (2000). What Should Comprehension Instruction Be the Instruction Of. In: P.B. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr, (Eds.). *Handbook of Reading Research, Volume 3*, pp.545-561. Mahwah, NJ: Lawrence Erlbaum.

- Rapp, D.N., Broek, P. van den, McMaster, K.L., Kendeou, P. & Espin, C.A. (2007). *Higher-order comprehension processes in struggling readers: a perspective for research and intervention*. *Scientific studies of reading*, 11 [4], pp. 289-312. Lawrence Erlbaum Associates, Inc.
- Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Newbury Park/London/New Delhi: Sage Publications.
- Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument*. Universiteit Twente, 1994.
- Van den Broek, P. (2012) Individual and developmental differences in reading comprehension: assessing cognitive processes and outcomes. In: J.P. Sabatini, E.R. Albro & T. O'Reilly (Eds.) (2012), *Measuring Up: Advances in How We Assess Reading Ability*. Landham, MD: Rowman & Littlefield Education.
- Van den Broek, P. & Espin, C.A. (2012). Connecting cognitive theory and assessment: measuring individual differences in reading comprehension. *School Psychology Review* 41 [3], pp. 315-325.
- Van den Broek, P. Lorch, R.F., Jr., Linderholm, T. & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition*, 29, pp. 1081-1087.
- 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., & Snow, C. (Eds.) (2001). *Literacy and motivation. Reading Engagement in Individuals and Groups*. Mahwah, NJ: Erlbaum.

Bijlagen

Bijlage 1 Uitwerking van referentieniveaus 1F en 2F voor Leesvaardigheid

Referentieniveaus 1F en 2F voor het lezen van zakelijke teksten

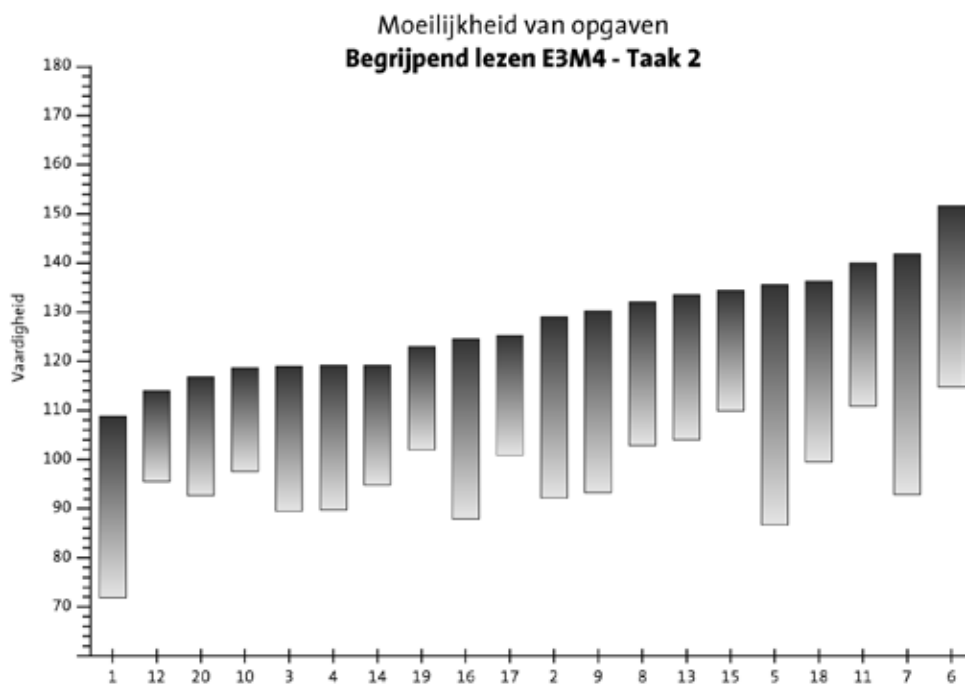
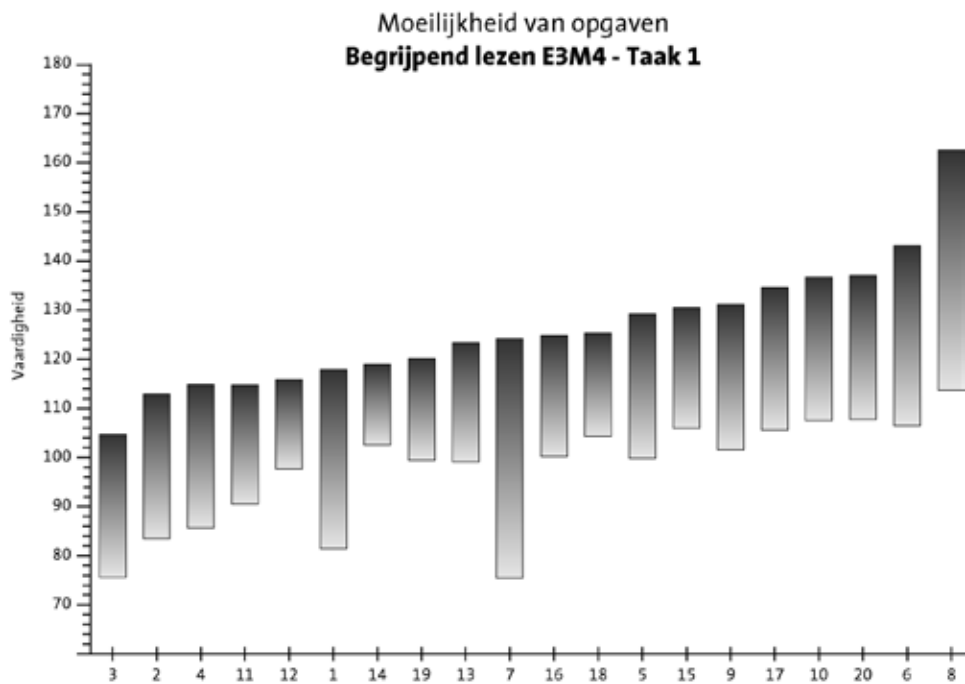
Leesvaardigheid Zakelijke teksten	Niveau 1F	Niveau 2F
Algemene omschrijving lezen zakelijke teksten	Kan eenvoudige teksten lezen over alledaagse onderwerpen en over onderwerpen die aansluiten bij de leefwereld.	Kan teksten lezen over alledaagse onderwerpen, onderwerpen die aansluiten bij de leefwereld van de leerling en over onderwerpen die verder van de leerling af staan.
Teksten:		
Tekstkenmerken	De teksten zijn eenvoudig van structuur; de informatie is herkenbaar geordend. De teksten hebben een lage informatie-dichtheid; belangrijke informatie is gemarkeerd of wordt herhaald. Er wordt niet te veel (nieuwe) informatie gelijktijdig geïntroduceerd. De teksten bestaan voornamelijk uit frequent gebruikte (of voor de leerlingen alledaagse) woorden.	De teksten zijn relatief complex, maar hebben een duidelijke opbouw die tot uiting kan komen in het gebruik van kopjes. De informatiedichtheid kan hoog zijn.
Taken:		
1. Lezen van informatieve teksten	Kan eenvoudige informatieve teksten lezen, zoals zaakvakteksten, naslagwerken, (eenvoudige) internetteksten, eenvoudige schematische overzichten.	Kan informatieve teksten lezen, waaronder schoolboek- en studieteksten (voor taal- en zaakvakken), standaardformulieren, populaire tijdschriften, teksten van internet, notities en schematische informatie (waarin verschillende dimensies gecombineerd worden) en het alledaagse nieuws in de krant.
2. Lezen van instructies	Kan eenvoudige instructieve teksten lezen, zoals (eenvoudige) routebeschrijvingen en aanwijzingen bij opdrachten (uit de methode).	Kan instructieve teksten lezen, zoals recepten, veelvoorkomende aanwijzingen en gebruiksaanwijzingen en bijsluiters van medicijnen.
3. Lezen van betogende teksten	Kan eenvoudige betogende teksten lezen, zoals voorkomend in schoolboeken voor taal- en zaakvakken, maar ook advertenties, reclames, huis- aan huisbladen.	Kan betogende, vaak redundante teksten lezen, zoals reclame-teksten, advertenties, folders, maar ook brochures van formele instanties of licht opiniërende artikelen uit tijdschriften.

Kenmerken van de taakuitvoering:		
Begrijpen	<p>Herkent specifieke informatie, wanneer naar één expliciet genoemde informatie-eenheid gevraagd wordt (letterlijk begrip).</p> <p>Kan (in het kader van het leesdoel) belangrijke informatie uit de tekst halen en kan zijn manier van lezen daar op afstemmen (bijvoorbeeld globaal, precies, selectief/gericht).</p>	<p>Kan de hoofdgedachte van een tekst weergeven en maakt onderscheid tussen hoofd- en bijzaken.</p> <p>Herkent beeldspraak (letterlijk en figuurlijk taalgebruik).</p> <p>Legt relaties tussen tekstdelen (inleiding, kern, slot) en teksten.</p> <p>Ordent informatie (bijvoorbeeld op basis van signaalwoorden) voor een beter begrip.</p>
Interpreteren	<p>Kan informatie en meningen interpreteren voor zover deze dicht bij de leerling staan.</p>	<p>Legt relaties tussen tekstuele informatie en meer algemene kennis.</p> <p>Kan de bedoeling van tekstgedeeltes en/of specifieke formuleringen duiden.</p> <p>Kan de bedoeling van de schrijver verwoorden.</p>
Evalueren	<p>Kan een oordeel over een tekst(deel) verwoorden.</p>	<p>Kan relaties tussen en binnen teksten evalueren en beoordelen.</p>

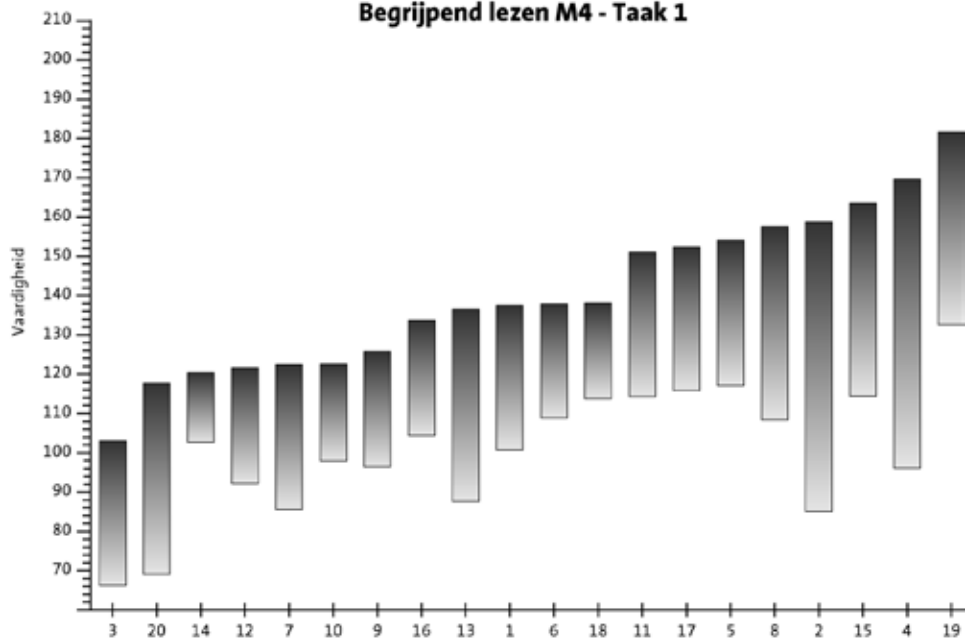
Referentieniveaus 1F en 2F voor het lezen van fictionele, narratieve en literaire teksten

Leesvaardigheid Fictionele, narratieve en literaire teksten	Niveau 1F	Niveau 2F
Algemene omschrijving Lezen fictionele, narratieve en literaire teksten	Kan jeugdliteratuur belevend lezen.	Kan eenvoudige adolescenten-literatuur herkendend lezen.
Teksten:		
	De structuur is eenvoudig. Het tempo waarin de spannende of dramatische gebeurtenissen elkaar opvolgen is hoog.	De structuur is helder. Het verhaal heeft een dramatische verhaallijn waarin de spanning af en toe wordt onderbroken door gedachten of beschrijvingen. Poëzie en liedjes hebben meestal een verhalende inhoud en emotionele lading.
Tekstkenmerken:		
Kenmerken van de taakuitvoering:	Herkent basale structuur-elementen, zoals wisselingen van tijd en plaats, rijm en versvorm. Kan meeleven met een personage en uitleggen hoe een personage zich voelt. Kan gedichten en verhaal-fragmenten parafraseren of samenvatten.	Herkent het genre. Herkent letterlijk en figuurlijk taalgebruik.
Interpreteren	Kan relaties leggen tussen de tekst en de werkelijkheid. Kan spannende, humoristische of dramatische passages in de tekst aanwijzen. Herkent verschillende emoties in de tekst, zoals verdriet, boosheid en blijdschap.	Kan bepalen in welke mate personages en gebeurtenissen herkenbaar en realistisch zijn. Kan personages typeren, zowel innerlijk als uiterlijk. Kan het onderwerp in de tekst benoemen.
Evalueren	Evalueert de tekst met emotieve argumenten. Kan met medeleerlingen leeservaringen uitwisselen. Kan interesse in bepaalde fictievormen aangeven.	Evalueert de tekst ook met realistische argumenten en kan persoonlijke reacties toelichten met voorbeelden uit de tekst. Kan met medeleerlingen leeservaringen uitwisselen en kan de interesse in bepaalde genres of onderwerpen motiveren.

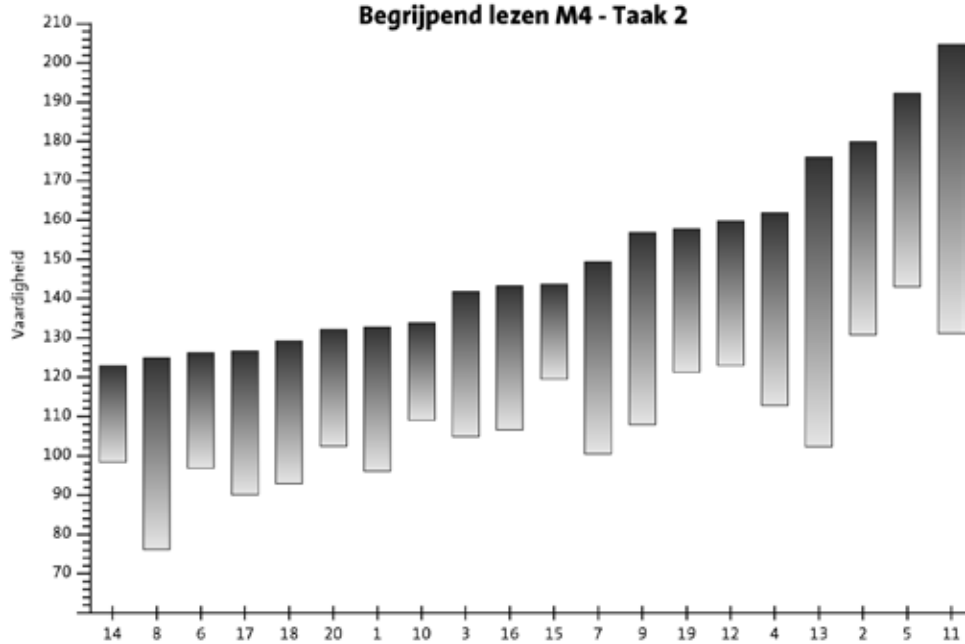
Bijlage 2 Moeilijkheid van de opgaven



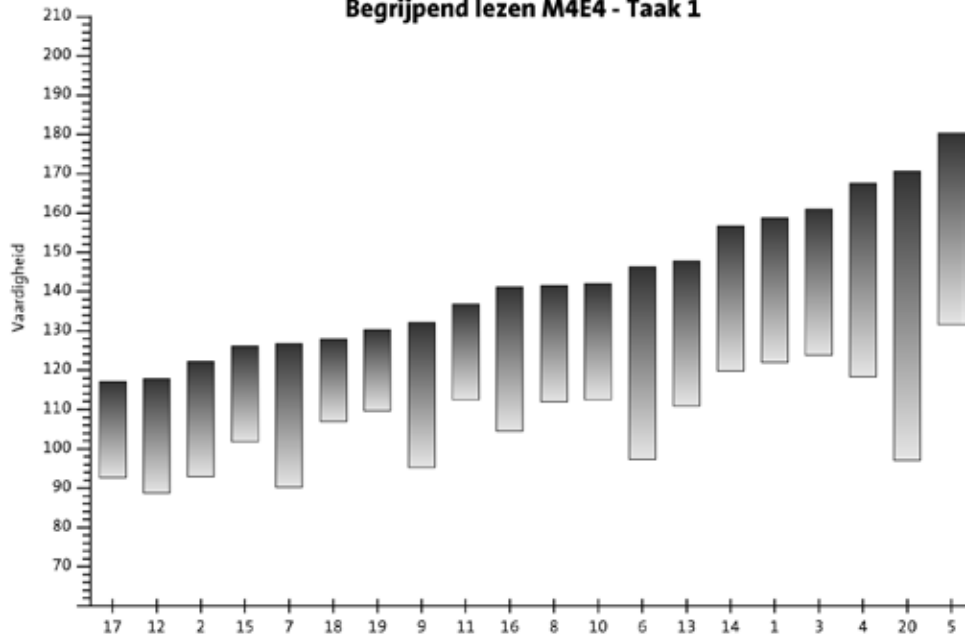
Moeilijkheid van opgaven
Begrijpend lezen M4 - Taak 1



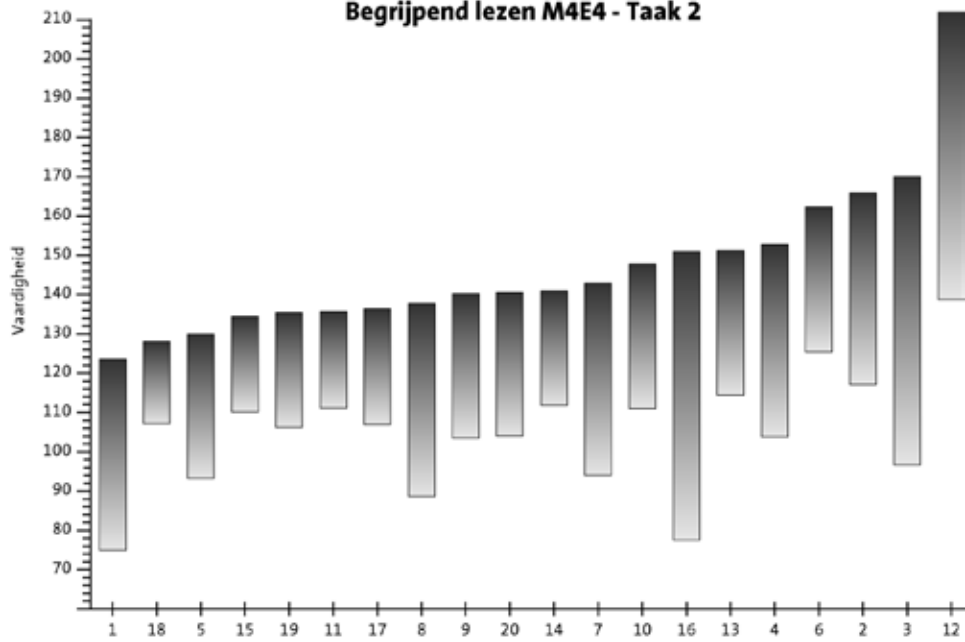
Moeilijkheid van opgaven
Begrijpend lezen M4 - Taak 2



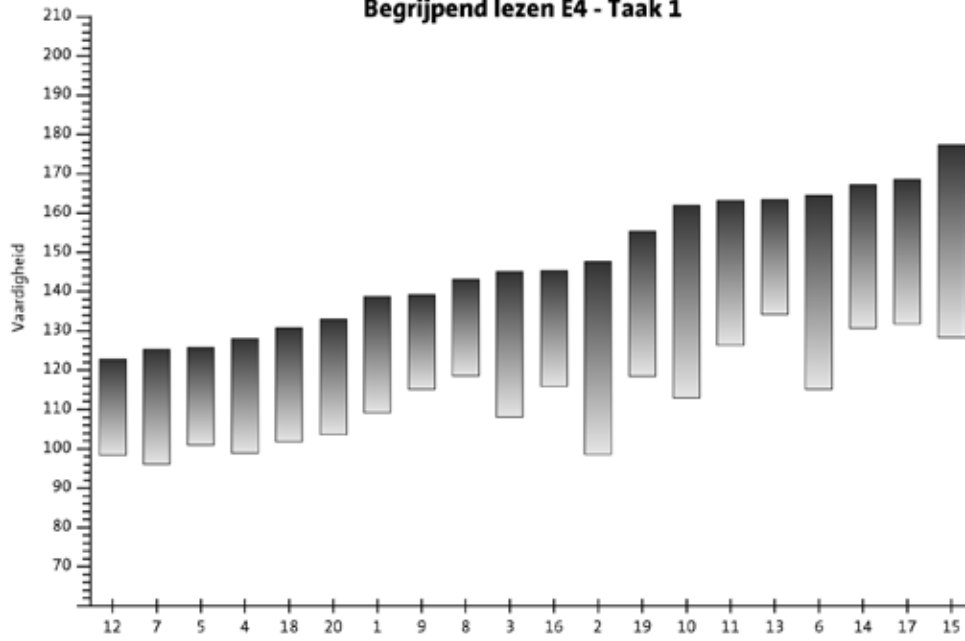
Moeilijkheid van opgaven
Begrijpend lezen M4E4 - Taak 1



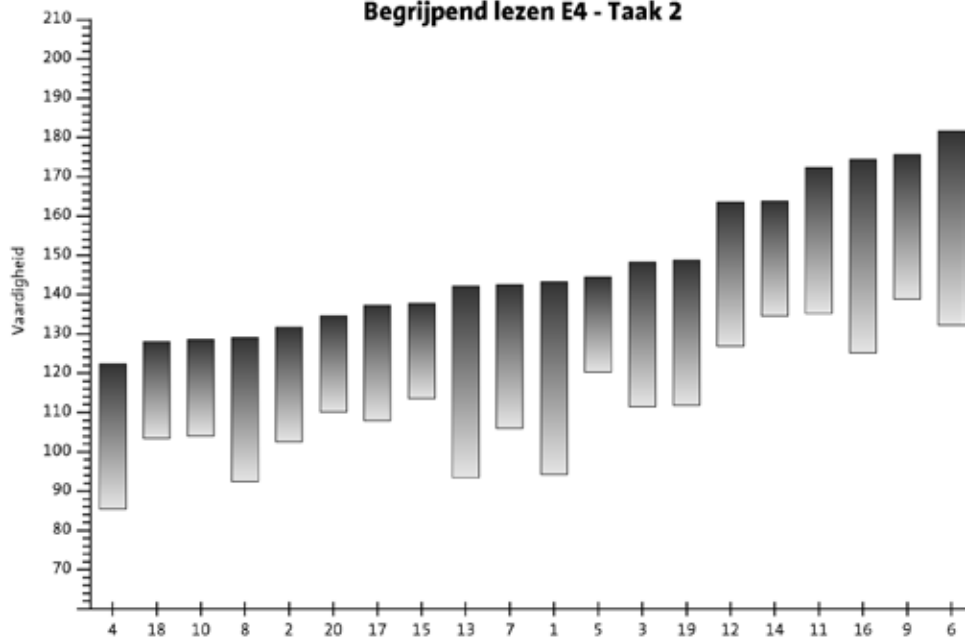
Moeilijkheid van opgaven
Begrijpend lezen M4E4 - Taak 2



Moeilijkheid van opgaven
Begrijpend lezen E4 - Taak 1



Moeilijkheid van opgaven
Begrijpend lezen E4 - Taak 2



Bijlage 3 **Klassieke en IRT-indices van de opgaven in de E3M4-toets**

Volgnr boekje	p-waarde	RIT	Beta	Info
1	0,84	0,36	-0,27	1,676
2	0,87	0,40	-0,25	2,082
3	0,90	0,37	-0,32	1,621
4	0,86	0,41	-0,23	2,207
5	0,78	0,46	-0,09	3,180
6	0,70	0,41	-0,03	2,677
7	0,82	0,30	-0,32	1,150
8	0,63	0,35	0,04	1,842
9	0,77	0,46	-0,08	3,297
10	0,73	0,47	-0,02	3,680
11	0,86	0,46	-0,18	2,929
12	0,85	0,55	-0,11	4,675
13	0,81	0,49	-0,10	3,773
14	0,83	0,59	-0,07	6,201
15	0,76	0,51	-0,03	4,493
16	0,80	0,45	-0,09	3,916
17	0,74	0,47	-0,04	3,554
18	0,79	0,55	-0,05	5,117
19	0,83	0,53	-0,10	4,410
20	0,72	0,47	-0,02	3,708
21	0,88	0,33	-0,36	1,327
22	0,79	0,39	-0,17	2,105
23	0,84	0,42	-0,19	2,469
24	0,84	0,42	-0,19	2,478
25	0,77	0,32	-0,22	1,368
26	0,65	0,42	0,05	2,963
27	0,74	0,33	-0,16	1,490
28	0,76	0,46	-0,07	3,359
29	0,78	0,39	-0,15	2,158
30	0,84	0,52	-0,11	4,171
31	0,70	0,48	0,01	3,894
32	0,87	0,54	-0,13	4,321
33	0,75	0,47	-0,05	3,460
34	0,84	0,48	-0,14	3,328
35	0,73	0,52	0,00	4,868
36	0,82	0,38	-0,21	1,923
37	0,80	0,50	-0,08	3,966
38	0,75	0,40	-0,10	2,398
39	0,81	0,54	-0,07	4,776
40	0,85	0,47	-0,16	3,113

Volgnr boekje	p-waarde	RIT	Beta	Info
1	0,74	0,40	-0,09	2,449
2	0,71	0,26	-0,23	0,767
3	0,90	0,30	-0,41	1,148
4	0,66	0,27	-0,13	0,828
5	0,63	0,43	0,07	3,037
6	0,72	0,47	-0,01	3,764
7	0,83	0,36	-0,23	1,840
8	0,66	0,35	-0,01	1,767
9	0,80	0,43	-0,13	2,928
10	0,82	0,47	-0,11	3,652
11	0,65	0,42	0,04	2,943
12	0,83	0,42	-0,17	2,641
13	0,77	0,32	-0,21	1,388
14	0,82	0,54	-0,07	5,512
15	0,62	0,36	0,05	1,854
16	0,75	0,46	-0,05	3,478
17	0,64	0,42	0,06	2,991
18	0,70	0,52	0,04	5,246
19	0,51	0,36	0,22	2,011
20	0,84	0,29	-0,38	1,028
21	0,77	0,39	-0,13	2,256
22	0,52	0,36	0,20	2,004
23	0,72	0,41	-0,05	2,615
24	0,63	0,35	0,03	1,831
25	0,44	0,35	0,31	2,017
26	0,80	0,44	-0,12	2,962
27	0,70	0,34	-0,09	1,635
28	0,82	0,30	-0,32	1,162
29	0,66	0,35	-0,02	1,758
30	0,73	0,51	0,00	4,805
31	0,51	0,28	0,20	0,946
32	0,58	0,43	0,13	3,185
33	0,64	0,27	-0,07	0,860
34	0,82	0,48	-0,11	3,704
35	0,65	0,53	0,09	5,711
36	0,70	0,41	-0,03	2,677
37	0,80	0,38	-0,19	2,010
38	0,79	0,38	-0,16	2,120
39	0,60	0,43	0,11	3,139
40	0,76	0,45	-0,07	3,370

Volgnr boekje	p-waarde	RIT	Beta	Info
1	0,63	0,42	0,12	2,626
2	0,85	0,40	-0,16	1,713
3	0,62	0,42	0,14	2,703
4	0,63	0,35	0,08	1,655
5	0,55	0,35	0,21	1,848
6	0,74	0,33	-0,12	1,262
7	0,83	0,36	-0,19	1,408
8	0,73	0,46	0,02	2,885
9	0,80	0,37	-0,14	1,598
10	0,73	0,46	0,03	2,926
11	0,75	0,50	0,03	3,503
12	0,87	0,38	-0,20	1,499
13	0,71	0,41	0,01	2,205
14	0,65	0,42	0,10	2,548
15	0,83	0,47	-0,08	2,524
16	0,75	0,39	-0,05	1,946
17	0,88	0,42	-0,16	1,824
18	0,81	0,52	-0,03	3,358
19	0,79	0,53	0,00	3,650
20	0,68	0,26	-0,12	0,738
21	0,84	0,29	-0,33	0,847
22	0,64	0,35	0,07	1,629
23	0,68	0,26	-0,12	0,735
24	0,71	0,34	-0,06	1,388
25	0,81	0,37	-0,16	1,517
26	0,60	0,42	0,15	2,756
27	0,76	0,32	-0,15	1,196
28	0,78	0,31	-0,20	1,097
29	0,76	0,39	-0,06	1,911
30	0,71	0,41	0,01	2,205
31	0,76	0,50	0,01	3,384
32	0,50	0,27	0,27	0,926
33	0,69	0,41	0,05	2,344
34	0,73	0,46	0,02	2,867
35	0,77	0,50	0,00	3,280
36	0,75	0,24	-0,30	0,614
37	0,77	0,45	-0,03	2,560
38	0,81	0,52	-0,02	3,378
39	0,77	0,45	-0,03	2,505
40	0,75	0,39	-0,05	1,931

Volgnr boekje	p-waarde	RIT	Beta	Info
1	0,75	0,33	0,00	2,710
2	0,74	0,33	-0,10	1,289
3	0,73	0,40	-0,01	2,098
4	0,82	0,41	-0,10	2,057
5	0,83	0,46	-0,08	2,490
6	0,65	0,35	0,05	1,601
7	0,83	0,41	-0,13	1,897
8	0,70	0,52	0,08	4,105
9	0,73	0,51	0,05	3,740
10	0,66	0,34	0,03	1,558
11	0,60	0,42	0,16	2,783
12	0,85	0,45	-0,11	2,246
13	0,55	0,48	0,23	4,239
14	0,56	0,42	0,20	2,921
15	0,57	0,35	0,18	1,807
16	0,70	0,47	0,06	3,152
17	0,56	0,42	0,21	2,955
18	0,80	0,43	-0,08	2,227
19	0,66	0,42	0,08	2,502
20	0,79	0,44	-0,06	2,349
21	0,76	0,32	-0,15	1,199
22	0,80	0,43	-0,07	2,271
23	0,71	0,41	0,02	2,219
24	0,85	0,34	-0,23	1,250
25	0,69	0,52	0,10	4,233
26	0,54	0,35	0,21	1,859
27	0,74	0,40	-0,04	2,002
28	0,82	0,36	-0,16	1,488
29	0,50	0,42	0,27	3,127
30	0,81	0,47	-0,05	2,723
31	0,53	0,42	0,24	3,049
32	0,60	0,42	0,16	2,793
33	0,76	0,32	-0,16	1,181
34	0,54	0,48	0,23	4,254
35	0,74	0,51	0,03	3,570
36	0,59	0,35	0,15	1,763
37	0,76	0,45	-0,02	2,606
38	0,82	0,47	-0,06	2,646
39	0,70	0,41	0,02	2,236
40	0,77	0,50	0,00	3,283

Cito helpt je inzicht te krijgen in je ontwikkeling en mogelijkheden. Door kennis, vaardigheden en competenties objectief meetbaar te maken en de ontwikkeling er van te volgen, kun je het beste uit jezelf halen, verantwoorde keuzes maken en beter richting geven aan je toekomst. Cito draagt daaraan bij door wereldwijd werk te maken van goed en eerlijk toetsen, vanuit de kernwaarden kundig, toonaangevend, integer, innovatief en betrokken.

Cito

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

Fotografie: Ron Steemers

