# Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 5

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

zeker weten

# Wetenschappelijke verantwoording

## Begrijpend lezen 3.0 voor groep 5

Marieke Tomesen
Anke Weekers
Maartje Hilte
Anke Jolink
Ron Engelen

Cito
Arnhem, 2016

# Inhoud

# 1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de LVS-toetsen Begrijpend lezen 3.0 voor groep 5. De toetsen Begrijpend lezen 3.0 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 basisonderwijs. Ze zijn ook geschikt voor toepassing in het speciaal basisonderwijs en het speciaal onderwijs cluster 2 en 4. Het betreft papieren toetsen[1] voor alle leerjaren.
In 2015 zijn de wetenschappelijke verantwoordingen voor de toetsen Begrijpend lezen 3.0 groep 3 en groep 4 uitgebracht. Te zijner tijd zullen ook de wetenschappelijke verantwoordingen met de gegevens van de (nog te verschijnen) toetsen Begrijpend lezen 3.0 voor de groepen 6 t/m 8 gefaseerd worden uitgebracht.

Deze verantwoording biedt, samen met de inhoud van het toetspakket Begrijpend lezen 3.0 voor groep 5, 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 3.0 groep 5 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 3.0 voor groep 5. 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 3.0 groep 5 van het Cito Volgsysteem primair en speciaal onderwijs beogen die vaardigheid te meten en de opgaven in deze toetsen zijn er dan ook operationaliseringen 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 5 van het Cito Volgsysteem primair en speciaal onderwijs zijn bestemd voor leerlingen in groep 5 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, met dien verstande 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 5 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.
Vanaf groep 5 wordt de groei die leerlingen bij begrijpend lezen doormaken kleiner dan in eerdere leerjaren het geval is. Het is daarom niet nodig om tweemaal in een schooljaar een toets Begrijpend lezen af te nemen. De leerkracht kan kiezen op welk van de twee afnamemomenten hij de leerlingen een toets laat maken. Kiest hij voor het medio-moment, dan maken de leerlingen in principe de toets M5 (want deze is genormeerd op het medio-moment). Kiest hij voor het einde-moment dan maken de leerlingen in principe de toets E5 (want deze is genormeerd op het einde-moment).
De toetsen Begrijpend lezen groep 5 kunnen ook gebruikt worden voor leerlingen in andere leerjaren die werken op het niveau van groep 5. 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 is een extra toets (E4M5) beschikbaar in het toetspakket voor groep 5. Door het toevoegen van een extra toets beslaan opeenvolgende toetsen kleinere leerstappen. De extra toets in deze uitgave is de toets voor het functioneringsniveau E4M5. Deze toets valt tussen de toetsen voor het functioneringsniveau E4 (einde groep 4) en M5 (medio groep 5). De toets voor functioneringsniveau E4M5 is de gemakkelijke variant van de toets M5. Aan een leerling die zich minder snel ontwikkelt in leesvaardigheid, kan medio groep 5 dus de toets E4M5 voorgelegd worden. Deze leerling hoeft zo niet een te moeilijke toets (M5) te maken, maar ook niet twee keer dezelfde toets (E4). Voor deze extra toets zijn de parameters van het reguliere afnamemoment M5 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 gemakkelijkere variant is van de opvolgende reguliere toets.

Vanwege de afvlakkende ontwikkelingscurve vanaf groep 5 zijn er vanaf M5 geen extra toetsen ('tussentoetsen') meer. Zoals hierboven vermeld, is er nog wel een E4M5-toets als gemakkelijke variant van de toets M5, maar er is geen M5E5-toets.

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 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 gemiddelde[2] 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 |
| **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 unidimensionale 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 wordt 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 vaardigheids-schaal. Dat geldt voor zowel individuele leerlingen als voor (gemiddelde) groepsresultaten. 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 vaardigheids-niveaus (die zijn afgebakend door percentielpunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk zoals "Koen heeft op afnamemoment medio leerjaar 5 vaardigheidsniveau IV behaald". Voor de leerkracht (en voor Koen en zijn ouders) bevat deze uitspraak waardevolle informatie.
De leerkracht kan op basis hiervan bijvoorbeeld besluiten om Koen 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 M5 had Koen vaardigheidsniveau IV en op tijdstip E5 was het vaardigheidsniveau III". Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 148, bijvoorbeeld, op tijdstip M5 en vaardigheidsscore 157 op tijdstip E5. 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 Koen 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: Koen 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 M5 behaalde Samira een vaardigheidsscore van 154 met een 67% betrouwbaarheidsinterval van 145-163. Bij de afname E5 behaalde Samira een vaardigheidsscore van 176; het bijbehorende betrouwbaarheidsinterval daarbij is 165-187. Aangezien de betrouwbaarheidsintervallen niet overlappen kunnen we zeggen dat Samira's 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. 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 vaardigheidsniveaus. 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 Volgsysteem 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 verwachtings-patroon 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 (Sijtstra, 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 gebruikgemaakt van methodes en wordt veel aandacht besteed aan het leren toepassen van leesstrategieën (Förrer & 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 boordelen 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 verwijswoorden 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). Het referentiekader onderscheidt voor taal vier domeinen: Mondelinge taalvaardigheid, Lezen, Schrijven en Begrippenlijst en taalverzorging. 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. De referentieniveaus Lezen 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.
Voor het leesonderwijs in het primair en speciaal onderwijs zijn vooral referentieniveaus 1F en 2F van belang. Referentieniveau 1F is een minimum- of drempelniveau dat ongeveer 75% van de leerlingen aan het eind van het primair en speciaal onderwijs kan halen. Het niveau 2F geldt als streefniveau voor het primair en speciaal onderwijs en wordt ook wel aangeduid als niveau 1S.
De niveaus 1F en 2F geven een eindpunt aan. Vanaf groep 3 zijn leerlingen onderweg naar het realiseren van de referentieniveaus. 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. Met onze toetsinhoud sluiten we aan bij de genoemde opbouw in leerstof over de leerjaren.

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 5: 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 gebruik gemaakt 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 zin 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 5 kan een toets maken die normaliter aan groep 4 wordt voorgelegd, en zijn vaardigheidsschatting kan behalve met de populatie van groep 5 ook vergeleken worden met de percentielen in de populatie van groep 4, met bijvoorbeeld de uitspraak: “De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 4”.
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 5 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).

#### 2.4.1.2 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 genoemd[3]. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 \mid \theta) \tag{2.1}$$

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

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

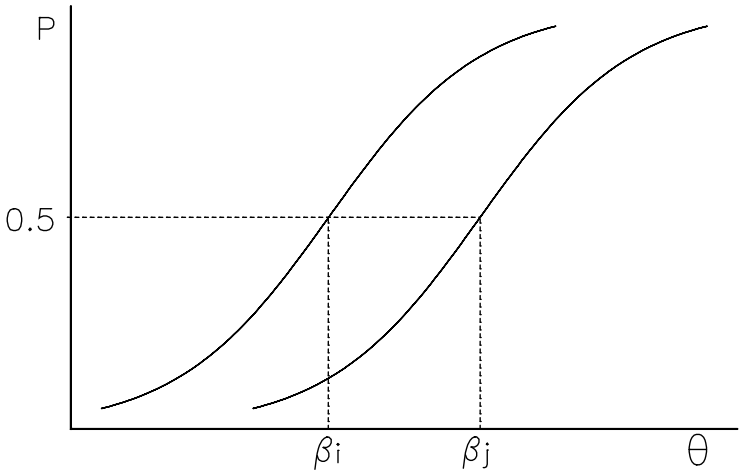
waarin $\beta_i$ de moeilijkheidsparameter van item *i* is. Dat is een onbekende grootheid die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.1 voor twee items, *i* en *j*, die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende 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 $\beta_i$, volgt

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

Daaruit volgt onmiddellijk een interpretatie voor de parameter $\beta_i$: 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 $\beta_i$ kan dus terecht omschreven worden als de moeilijkheidsparameter van item *i*. De implicatie van het bovenstaande is dat 'moeilijkheid' en 'vaardigheid' op dezelfde schaal liggen.

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

*Figuur 2.1 Twee itemresponscurven in het Raschmodel*




Formule (2.2) is geen beschrijving van de werkelijkheid, het is een hypothese over de werkelijkheid die getoetst kan worden op haar houdbaarheid. Hoe zo'n toetsing grofweg verloopt, is te verduidelijken aan de hand van figuur 2.1. Daaruit blijkt dat, voor welk vaardigheidsniveau dan ook, de kans om 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 steekproefgrootheid ('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).

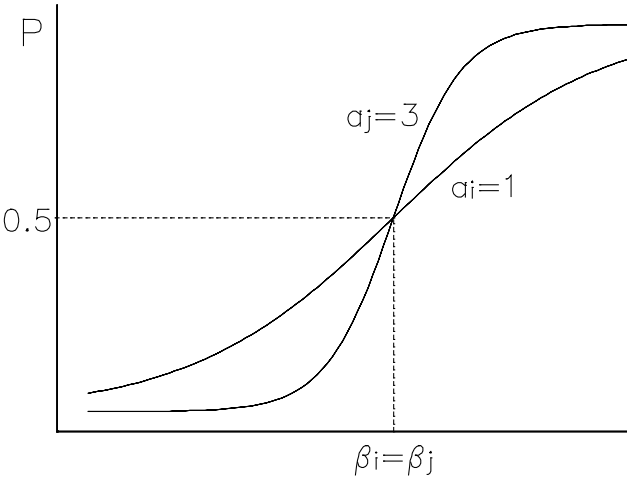
[4] Een gedetailleerde uiteenzetting hierover kan men vinden in Verhelst, 1992.

De itemresponsfunctie van het OPLM is gegeven door

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

waarin $a_i$ de zogenaamde discriminatie-index van het item is. Door deze indices te beperken tot (positieve) gehele getallen, en door ze a-priori als constanten in te voeren, is het mogelijk CML-schattingen van de itemparameters $\beta_i$ 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*




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-schattingsmethode 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 gebruik gemaakt 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 $\theta$-parameter (de vaardigheidsscore) zoals die gedefinieerd is in bijvoorbeeld het OPLM. Er wordt namelijk niet één enkele puntschatting voor $\theta$ bepaald, maar er wordt een reeks van mogelijke waarden voor $\theta$ geschat die elk een bepaalde kans hebben om geobserveerd te worden. *Plausible values* zijn random trekkingen uit deze (geschatte) verdeling voor de vaardigheid $\theta$ 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 $\theta$ zou tot bias in sommige populatieparameters leiden. Zo zou de variantie bij gebruik van de ML- of WML-schatting van $\theta$ bijvoorbeeld overschat worden.

# 3 Beschrijving van de toets

## 3.1 Opbouw en structuur van de toetsen

Het toetspakket Begrijpend lezen 3.0 voor groep 5 uit het Cito Volsysteem primair en speciaal onderwijs bevat in totaal drie papieren toetsen: E4M5, M5 en E5.
De toetsen kunnen in groep 5 op één of twee afnamemomenten afgenomen worden: halverwege het schooljaar (half januari/half februari) en/of aan het einde van het schooljaar (juni). De toetsen M5 en E5 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 een extra toets (E4M5) voor leerlingen met een vertraagde ontwikkeling. Deze toets is de gemakkelijke variant van de toets M5 en kan voorgelegd worden aan leerlingen voor wie de toets M5 nog net te moeilijk is. Deze leerlingen hoeven dan niet nogmaals de toets E4 te maken. Ook deze extra toets is bedoeld voor afname op de reguliere afnamemomenten.

Vanaf groep 5 wordt de groei die leerlingen bij begrijpend lezen doormaken kleiner dan in de leerjaren daarvoor. Het is daarom niet nodig om tweemaal in een schooljaar een toets Begrijpend lezen af te nemen. De leerkracht kan kiezen op welk van de twee afnamemomenten hij de leerlingen een toets laat maken. Die afvlakkende ontwikkelingscurve is ook de reden dat er vanaf M5 geen extra toetsen meer zijn. Er is nog wel een E4M5-toets als gemakkelijke variant van de toets M5, maar er is geen M5E5-toets.

*Opbouw*
De toetsen Begrijpend lezen 3.0 voor groep 5 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 25 opgaven behorende bij een aantal teksten. In totaal bevat elke toets dus 50 opgaven.

*Vorm*
De toetsen voor groep 5 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 antwoord-alternatieven.

*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 M5-toets voor leerlingen medio groep 5) 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 M5-leerling maakt een toets E4) of een volgend afnamemoment (een M5-leerling maakt de toets E5).

*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.
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 concentratie-problemen, 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 Computer-programma 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 E5 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling overeenkomt met de score van de gemiddelde leerling eind groep 5. 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 fases hieronder nader toelichten.

Deze informatie vormt een aanvulling op de inhoudsverantwoording die opgenomen is in het toetspakket Begrijpend lezen 3.0 voor groep 5 (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 3.0 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.

Terwijl voor de toetsen van groep 4 het streven nog was dat de helft van de teksten informeel zou zijn en de andere helft formeel (zakelijk), was voor de toetsen van groep 5 het streven dat meer dan de helft (60-70%) van de teksten formeel zou zijn en minder dan de helft (30-40%) informeel. Net als bij de toetsen van groep 4, zouden de formele teksten voor het grootste deel uit informatieve teksten moeten bestaan. Daarnaast zouden instructieve en betogende teksten opgenomen moeten worden. De informele teksten zouden uit verhalende teksten bestaan en een poëzietekst. Voor de tekstgenres was het streven dat er zoveel mogelijk diversiteit is.

*Indeling in opgavenvormen*
De opgaven in de toetsen Begrijpend lezen 3.0 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 van groep 5 was het streven dat er 25% openplaatsopgaven opgenomen werden, 65% tekstvragen en 10% voorspelopgaven.

*Indeling in vaardigheden*
Bij het construeren van de opgaven onderscheidden 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 kunnen opgaven nader worden gekarakteriseerd naar hun aard. 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.
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 tijdsperioden, (hoofd)personen.
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.
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.

*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).
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.
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.

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.

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 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. Hierbij moet benadrukt worden dat in werkelijkheid de vaardigheden Begrijpen en Interpreteren niet zo duidelijk van elkaar te scheiden zijn en dat ze niet opgevat kunnen worden als te isoleren vaardigheden van begrijpend lezen (zie ook hieronder). Voor de selectie van de uiteindelijke opgaven spelen ze dan ook een minder belangrijke rol dan de tekstsoorten en opgavenvormen. Bij het samenstellen van de taken Begrijpend lezen zijn we ervan uitgegaan dat de verwerkingsprocessen ‘begrijpen van geschreven teksten’ (kortweg Begrijpen) en ‘interpreteren van gesproken teksten’ (kortweg: Interpreteren) beide een belangrijke rol in het leesproces vervullen. Echter, het interpreteren van geschreven teksten neemt in de bovenbouw van het basisonderwijs een steeds belangrijker plaats in.

*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 wélke 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 opgaven-vormen, 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. Bij de opzet van de proeftoetsingen is ervoor gekozen om een taak met nieuwe opgaven te koppelen aan de reguliere LVS-afname Begrijpend lezen. Zo viel het voor de leerlingen niet op dat ze aan een proeftoets deelnamen en konden de leerkrachten gewoon de ontwikkeling van hun leerlingen blijven volgen.

Bij proeftoetsingen is in 2013 halverwege en aan het einde van leerjaar 5 een aantal opgaven voorgelegd aan leerlingen. Daarbij zijn op het afnamemoment medio groep 5 265 nieuwe opgaven voorgelegd aan 1910 leerlingen van groep 5 verdeeld over 14 boekjes. Elke deelnemende school maakte op het medio-moment één taak met 24-26 nieuwe opgaven, naast de LVS-toets Begrijpend lezen van de tweede generatie. De nieuwe opgaven kwamen in een of twee boekjes voor en werden gemiddeld door 182 leerlingen gemaakt.

Op het afnamemoment einde groep 5 zijn 171 items voorgelegd aan 2220 leerlingen van groep 5 verdeeld over 6 boekjes. Elk boekje bestond uit 50 opgaven verdeeld over 2 taken. Beide taken betroffen nieuw geconstrueerde opgaven. De opgaven kwamen in een of twee boekjes voor en werden gemiddeld door 649 leerlingen gemaakt. Op het eind-moment kon de proeftoets niet gekoppeld worden aan de afname van de tweede generatie, omdat de toetsen Begrijpend lezen tweede generatie geen einde-moment kennen vanaf groep 5.

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 vaardige leerlingen van de minder vaardige leerlingen te kunnen onderscheiden. Ook vonden de praktijkdeskundigen bepaalde teksten aan de lange kant. Voor een goede afwisseling in lengte van teksten, zijn sommige van die lange teksten niettemin wel opgenomen.

*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 de normeringsonderzoeken leken dus al zoveel mogelijk op de definitieve taken. Daarnaast werden in de normeringsonderzoeken reserveteksten met bijbehorende opgaven meegenomen voor het geval er onverhoopt toch nog teksten en opgaven zouden uitvallen vanwege slecht functioneren. Op die manier konden zo nodig hele teksten vervangen worden door andere.

*Samenstelling definitieve toetsen*
Na het normeringsonderzoek is van alle opgaven opnieuw de p-waarde en de rir bepaald. 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, de inhoud (het onderwerp) en de lengte van de teksten en de opgavenvorm. Daarnaast is ervoor gezorgd dat het aantal opgaven in verhouding staat tot de lengte van de tekst waarbij de opgaven horen.

Hoewel de afweging van inhoudelijke en psychometrische criteria bij toetsen Begrijpend lezen extra lastig is (er zijn immers niet alleen overwegingen op het niveau van individuele opgaven, maar ook op het niveau van de tekst), kon in de meeste gevallen aan zowel de psychometrische (p-waarde, $r_{ir}$) als de inhoudelijke criteria voldaan worden. In enkele gevallen zijn er op basis van de inhoudelijke criteria opgaven opgenomen die (net) niet de gewenste psychometrische waarden hadden.

Enkele opgaven vielen af, omdat die in het normeringsonderzoek een te hoge of te lage moeilijkheid of een te laag discriminerend vermogen vertoonden. Een tekst - met psychometrisch goede items - viel af omdat er van scholen erg veel inhoudelijke kritiek kwam op die tekst. Deze teksten met opgaven zijn vervangen door reserveteksten met bijbehorende opgaven. 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.

Bij de samenstelling van de toetsen is rekening gehouden met een mogelijk volgorde-effect: de volgorde van de opgaven in de definitieve toetsen wijkt nauwelijks af van die van het normeringsonderzoek.

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 5 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 M5 ligt het merendeel van de balkjes op en rond de gemiddelde vaardigheidsscore behorende bij medio groep 5. Er is een goede spreiding in moeilijkheidsgraad van de opgaven: er zijn makkelijke opgaven (die liggen onder de lijn van M5), opgaven van gemiddelde moeilijkheid (doorkruisen de lijn van M5) en moeilijke opgaven (liggen boven de lijn van M5). Bij deze toets kunnen ook de betere leerlingen laten zien wat ze in huis hebben. Bij de gemakkelijke variant van de toets M5, de toets E4M5, liggen vrijwel alle balkjes ónder de lijn van M5 (dus onder de gemiddelde vaardigheidsscore). Uit de figuren van E4M5 is duidelijk op te maken dat deze toets over de gehele linie erg makkelijk is en dus geschikt voor de zwakkere lezers. Bij de toets E5 liggen de opgaven veelal iets hoger op de vaardigheidsschaal, rond de gemiddelde vaardigheidsscore eind groep 5. Net als de toets M5 bevat deze toets makkelijke opgaven (die liggen onder de stippellijn van E5), opgaven van gemiddelde moeilijkheid (doorkruisen de lijn van E5) en moeilijke opgaven (liggen boven de lijn van E5). Ook is te zien dat de groei in vaardigheid tussen de momenten M5 en E5 maar klein is: de lijntjes van de gemiddelde vaardigheidsscores M5 resp. E5 liggen vrij dicht op elkaar.

In alle figuren is zichtbaar dat de toetsen relatief veel gemakkelijke opgaven bevatten, oftewel opgaven behorend bij lagere vaardigheidsniveaus dan het gemiddelde vaardigheidsniveau 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 een zo groot mogelijk variatie in vaardigheidsniveaus enerzijds en het creëren van de voorwaarden voor een prettige toetservaring voor leerling en leerkrachten anderzijds.

### 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 5 voor meer dan de helft (60-70%) uit formele (zakelijke) teksten zou bestaan en minder dan de helft (30-40%) uit informele teksten. De formele teksten zouden voor het grootste deel uit informatieve teksten moeten bestaan. Daarnaast zouden instructieve en betogende teksten opgenomen moeten worden. De informele teksten zouden uit verhalende teksten bestaan en een poëzietekst.

In tabel 3.1 is te zien dat bovenstaande redelijk opgaat voor de toetsen van groep 5. Het aandeel formele teksten varieert tussen 64% en 72%. In de toetsen E4M5 en M5 bestaan de formele teksten inderdaad voor het grootste deel uit informatieve teksten. In de toets E5 zijn naar verhouding veel betogende teksten opgenomen. Het is onmogelijk om precies de gewenste verdelingen te realiseren in de uiteindelijke toetsen. Daarvoor spelen bij de selectie van opgaven teveel overwegingen tegelijkertijd een rol. Dat geldt overigens ook voor de andere verdelingen die in deze paragraaf worden besproken. Naast de tekstsoorten, moesten 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 5: percentage teksten (aantal teksten)*

| Tekstsoort / Toets | formeel | | | informeel | | Totaal aantal teksten |
|---|---|---|---|---|---|---|
| | Informatief | Instructief | Betogend | Verhalend | Poëzie | |
| **E4M5** | 32% (6) | 16% (3) | 16% (3) | 32% (6) | 5% (1) | 19 |
| **M5** | 33% (6) | 28% (5) | 11% (2) | 22% (4) | 6% (1) | 18 |
| **E5** | 25% (4) | 13% (2) | 31% (5) | 25% (4) | 6% (1) | 16 |

*Verdeling van tekstgenres*

Voor de tekstgenres was het streven om zoveel mogelijk diversiteit te creëren. In tabel 3.2 is te zien dat alle toetsen in groep 5 uit 9 of 10 verschillende tekstgenres bestaan. Hiermee zijn de toetsen heel gevarieerd geworden. Sommige teksten bleken overigens in te delen bij meerdere genres: zo waren er artikelen waar ook instructies/aanwijzingen in werden gegeven of een verhaal waar ook een recept in stond. In zulke gevallen zijn de teksten ingedeeld bij het genre waarmee ze de meeste kenmerken gemeen hadden.

*Tabel 3.2 Tekstgenres in Begrijpend lezen groep 5: percentage teksten (aantal teksten)*

| Toets / Tekstgenre | E4M5 | M5 | E5 |
|---|---|---|---|
| **Artikel** | 32% (6) | 28% (5) | 13% (2) |
| **(Nieuws)bericht** | | 6% (1) | |
| **Instructie** | 5% (1) | 22% (4) | 6% (1) |
| **Speluitleg** | 11% (2) | | 6% (1) |
| **Recept** | | | 6% (1) |
| **Advies** | | 6% (1) | |
| **Reclame** | 5% (1) | 6% (1) | 6% (1) |
| **Oproep** | 5% (1) | | 19% (3) |
| **Recensie** | 5% (1) | 6% (1) | 6% (1) |
| **Brief** | 5% (1) | 6% (1) | 6% (1) |
| **Verhaal** | 26% (5) | 17% (3) | 25% (4) |
| **Gedicht** | 5% (1) | 6% (1) | 6% (1) |
| **Totaal aantal teksten** | 19 | 18 | 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 van groep 5 was het streven 25% openplaatsopgaven, 65% tekstvragen en 10% voorspelopgaven op te nemen. In tabel 3.3 is te zien dat deze streefverdelingen bij benadering gehaald zijn. Het aandeel openplaatsopgaven in de toetsen M5 en E5 is iets groter dan vooraf gewenst en het aandeel tekstvragen iets kleiner dan gewenst. Gezien de vele beperkende randvoorwaarden is de verdeling van opgavenvormen goed gelukt.

*Tabel 3.3 Opgavenvormen in Begrijpend lezen groep 5: percentage opgaven (aantal opgaven)*

| Opgavenvorm / Toets | Openplaats-opgave | Vraag over tekst | Voorspel-opgave | Totaal aantal opgaven |
|---|---|---|---|---|
| **E4M5** | 22% (11) | 68% (34) | 10% (5) | 50 |
| **M5** | 30% (15) | 60% (30) | 10% (5) | 50 |
| **E5** | 28% (14) | 62% (31) | 10% (5) | 50 |

*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 gedekt zijn in een toets.

Door de onderscheiden opgavenvormen en onderliggende inhoudsaspecten te verdelen over de vaardigheden Begrijpen en Interpreteren, hebben we ons er tijdens de constructiefase van verzekerd dat de vaardigheid in begrijpend lezen in al haar facetten en van alle kanten belicht werd. Het bleek in deze fase niet mogelijk om bij elke tekst alle beschikbare categorieën in te zetten, omdat niet alle teksten zich daar even goed voor lenen. Bovendien zijn in de fase van proeftoetsing teksten en opgaven uitgevallen.

Tabel 3.4 geeft de uiteindelijke verdeling weer van de opgaven over de onderscheiden vaardigheden Begrijpen en Interpreteren. Er zijn zowel opgaven opgenomen die een beroep doen op de vaardigheid Begrijpen als opgaven die een beroep doen op de vaardigheid Interpreteren. Zoals hieronder te zien is in tabel 3.4, is het percentage opgaven dat vooral een beroep doet op Interpreteren – zoals beoogd – in de meerderheid en neemt dit percentage toe naarmate de toets een hogere moeilijkheidsgraad heeft.

De exacte verdeling van inhoudsaspecten over vaardigheidsaspecten is niet belangrijk, deze werd voornamelijk bepaald door de aard van de teksten en de aard van de opgaven die na proeftoetsing op psychometrische gronden konden worden behouden. Hierbij moet in gedachten gehouden worden dat de vaardigheden Begrijpen en Interpreteren én de diverse inhoudsaspecten in werkelijkheid niet zo duidelijk van elkaar te scheiden zijn (vgl. paragraaf 3.2.1). De vaardigheden Begrijpen en Interpreteren liggen in elkaars verlengde en opgaven meten niet zelden zowel Begrijpen als Interpreteren. We kunnen ze dan ook niet opvatten als te isoleren vaardigheden en aspecten van begrijpend lezen. Op basis van intern onderzoek weten we dat de correlatie tussen begrijpen en interpreteren 0,98 is. Ook dit bevestigt het feit dat de opgaven op één vaardigheidsschaal liggen.

*Tabel 3.4 Vaardigheden in Begrijpend lezen groep 5: percentage opgaven (aantal opgaven)*

| Vaardigheid / Toets | B | I | Totaal aantal opgaven |
|---|---|---|---|
| **E4M5** | 36% (18) | 64% (32) | 50 |
| **M5** | 18% (9) | 82% (41) | 50 |
| **E5** | 12% (6) | 88% (44) | 50 |

### 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 E4M5, M5 en E5. Opgemerkt dient te worden dat de toets E4M5 als gemakkelijke versie voor moment M5 wordt gebruikt. De toets E4M5 is afgenomen in zowel de normeringsonderzoeken E4 als M5 (zie ook design hoofdstuk 4). In tabel 3.5 worden de beschrijvende gegevens van de toetsen E4M5, M5 en E5 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 E4M5 en M5 gebaseerd op 3547 leerlingen en voor E5 op 1724 leerlingen.

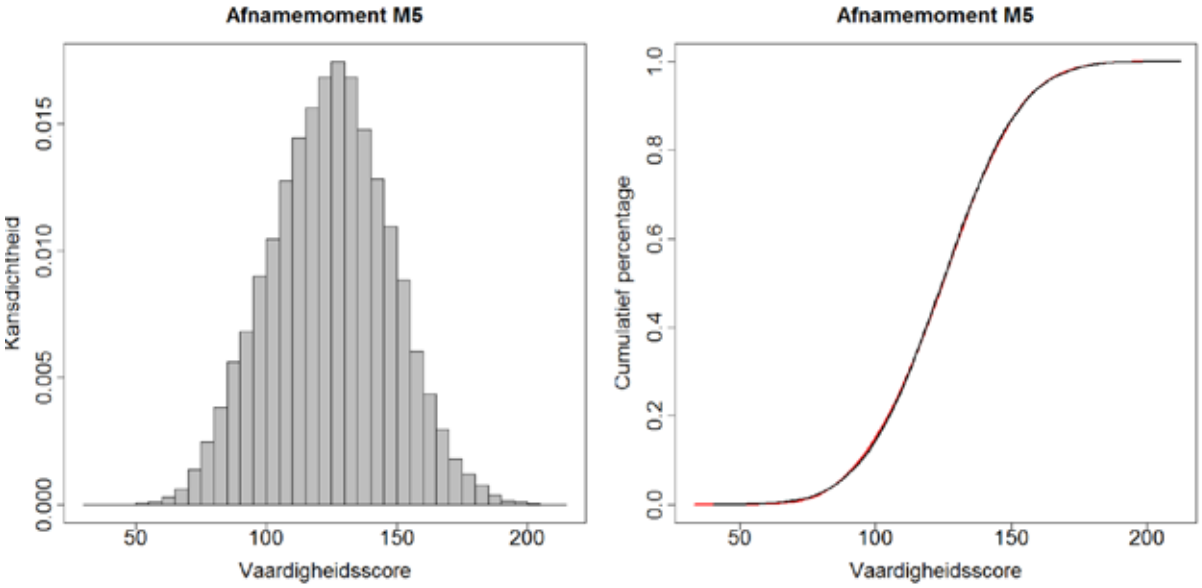
*Tabel 3.5 Beschrijvende gegevens toetsen E4M5, M5 en E5 op ruwe score en vaardigheidsschaal*

| Score | Gemiddelde | Standaarddeviatie | Kurtosis | Scheefheid |
|---|---|---|---|---|
| E4M5 Ruwe score | 37.9 | 8.45 | .716 | -1.013 |
| E4M5 Vaardigheid | 155.02 | 26.20 | -.183 | -.002 |
| M5 Ruwe score | 35.5 | 8.66 | .161 | -.787 |
| M5 Vaardigheid | 155.02 | 26.20 | -.183 | -.002 |
| E5 Ruwe score | 34.4 | 8.54 | -.041 | -.660 |
| E5 Vaardigheid | 159.51 | 24.57 | -.207 | .062 |

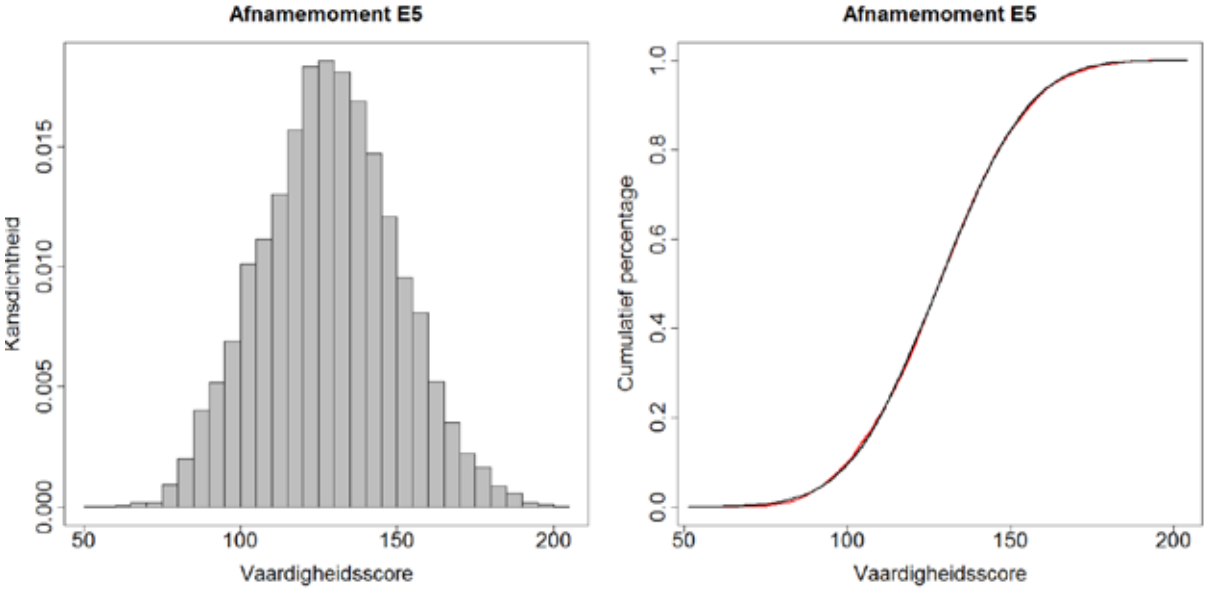
De ruwe scores zijn licht scheef verdeeld. Dit is niet verwonderlijk. De toetsen worden samengesteld rond een verwachte (gewenste) p-waarde van .70 voor de reguliere versies M5 en E5 en een verwachte (gewenste) p-waarde van .75-.80 voor de toets E4M5 als makkelijkere versie voor de toets M5.
De vaardigheidsverdelingen, de scores die gebruikt worden om leerlingen te vergelijken en te volgen, zijn 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 5 terwijl de gemiddelden en standaarddeviaties in tabel 3.3 op de (getransformeerde) overkoepelende vaardigheidsschaal worden weergegeven (zie voor een toelichting hoofdstuk 4).

*Figuur 3.1 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling afnamemoment M5*




*Figuur 3.2 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling afnamemoment E5*



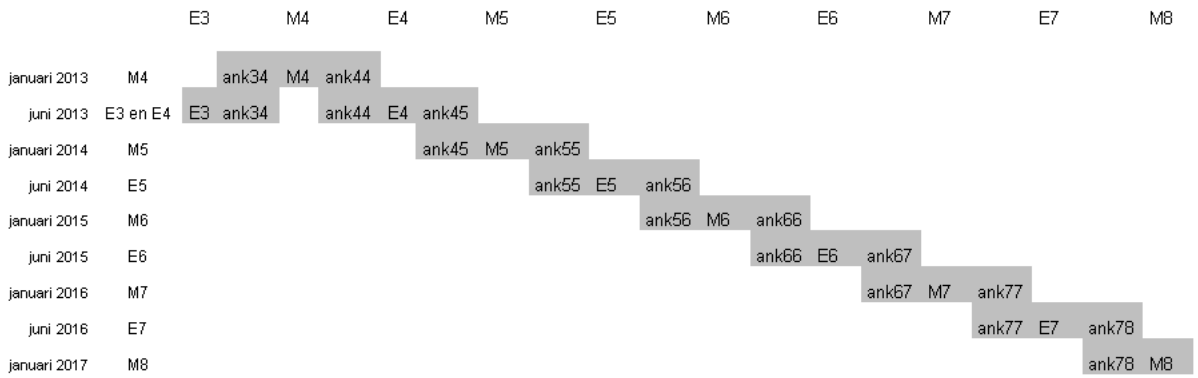
# 4 Kalibratie en normering

## 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 afnamemomenten, bijvoorbeeld E4 en E5. Met andere woorden, het dataverzamelings-design, 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. Voor E5 betekent dit dat de leerlingen in het kalibratie- en normeringsonderzoek items krijgen voorgelegd die specifiek voor E5 zijn geconstrueerd, en een (minderheid aan) items die geconstrueerd zijn voor M5 en M6.
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 ank45 de groep items aan die enerzijds bestaat uit items geconstrueerd voor E4 en anderzijds uit items geconstrueerd voor M5. Die items zijn dus zowel eind groep 4 als medio groep 5 afgenomen. De groep items ank55 bevat items voor M5 en E5, die dus zowel medio groep 5 als eind groep

5 zijn afgenomen en de groep items ank56 bevat items voor E5 en M6. 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 5 met afnamemomenten M5 en E5 vindt de kalibratie plaats op basis van afname-momenten E4, M5, E5 en M6. Dit sluit aan bij de inhoudelijke kenmerken van de aangeboden opgaven, een sterke leerling in groep 5 zal wel opgaven uit groep 6 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 5 betreft het dus het gedeelte van het macrodesign dat in figuur 4.2 hieronder is weergegeven.

*Figuur 4.2 Gedeelte macrodesign waarop kalibratie leerjaar 5 is gebaseerd*

| | | **E4** | | **M5** | | **E5** | | **M6** | |
|---|---|---|---|---|---|---|---|---|---|
| Juni 2013 | ank44 | E4 | ank45 | | | | | | |
| Januari 2014 | | | ank45 | M5 | ank55 | | | | |
| Juni 2014 | | | | | ank55 | E5 | ank56 | | |
| Januari 2015 | | | | | | | ank56 | M6 | ank66 |

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 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 kalibraties. 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 toets Begrijpend lezen M5 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 op afnamemoment M5 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 M5 aangevuld kan worden met resultaten uit dataretour van de tweede generatie LVS-toetsen (zie Keuning et al., 2015). Voor afnamemoment E5 bestaat er geen toets Begrijpend lezen LVS 2.0. Op dit afnamemoment maakten leerlingen uitsluitend twee taken met nieuw materiaal.

In figuur 4.3 en 4.4 worden de *embedded field* designs weergegeven voor respectievelijk toetsversie M5 en toetsversie E5. In het design kunnen we zien dat er zeven toetsversies M5 en vijf toetsversies E5 zijn afgenomen. Op afnamemoment M5 maakten leerlingen volgens het design een taak uit de M5-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 tot en met afnamemoment M5 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. Vanaf meetmoment M5 is de vooruitgang die de gemiddelde leerling boekt op begrijpend lezen te klein om tussentoetsen uit te kunnen geven. Toetsen voor opvolgende afnamemomenten volstaan dan. De taken E4M5 deel 1 en E4M5 deel 2 vormen tezamen de beoogde uitgave E4M5 LVS 3.0 en zijn ook eind groep 4 afgenomen. De taken M5 deel 1 en M5 deel 2 vormen tezamen de beoogde uitgave M5 LVS 3.0. De taken E5 deel 1 en E5 deel 2 vormen tezamen de beoogde uitgave E5 LVS 3.0. De taken M5 anker, E5 anker en M6 anker in de designs vormen opgaven uit de beoogde selectie van het betreffende afnamemoment.

Voor zowel E4M5, M5, als E5 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 E4M5, M5 en E5.

*Figuur 4.3 Design LVS 3.0 M5*

| Toets-versie | M5 LVS 2.0 | E4M5 deel 1 | E4M5 deel 2 | E4M5 reserve | M5 deel 1 | M5 deel 2 | M5 reserve | E5 anker | Leerlingen |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ■ | ■ | ■ | | | | | | 246 |
| 2 | ■ | | | | ■ | ■ | | | 260 |
| 3 | ■ | | ■ | | | | | ■ | 264 |
| 4 | ■ | ■ | | ■ | | | | | 267 |
| 5 | ■ | | | | ■ | | ■ | | 257 |
| 6 | ■ | | | | | ■ | | ■ | 260 |
| 7 | ■ | | | ■ | | | ■ | | 213 |

*Figuur 4.4 Design LVS 3.0 E5*

| Toets-versie | M5 anker | E5 deel 1 | E5 deel 2 | E5 reserve | M6 anker | Leerlingen |
|---|---|---|---|---|---|---|
| 1 | ■ | ■ | | | | 370 |
| 2 | | ■ | ■ | | | 360 |
| 3 | ■ | | | ■ | | 367 |
| 4 | | | ■ | | ■ | 350 |
| 5 | | | | ■ | ■ | 338 |

Zoals te zien in het design vormt het deel M5 LVS 2.0 een stevig anker tussen de toetsboekjes voor afnamemoment M5. Zowel binnen afnamemoment M5 als binnen afnamemoment E5 werd er geankerd door middel van de taken met nieuw ontwikkeld materiaal voor LVS 3.0. Tussen de beoogde reguliere toets M5 en de aangrenzende tussentoets E4M5 voor LVS 3.0 is geankerd in normeringsonderzoek M5 LVS 3.0. De tussentoets E4M5 werd meegenomen in normeringsonderzoek E4 en M5. Hierdoor is er ook een ankering over de afnamemomenten E4 en M5 heen. Het normeringsonderzoek E4 is verantwoord in de wetenschappelijke verantwoording voor Begrijpend lezen 3.0 groep 4. Voor de ankering tussen afnamemomenten M5 en E5 en E5 en M6, waar geen tussentoetsen meer worden uitgegeven, werd gezorgd door items voor de beoogde uitgave van het opvolgende dan wel voorgaande afnamemoment op te nemen.

De leerlingen op afnamemoment M5 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 M5. 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). Bij de normering van afnamemoment E5 was dit niet mogelijk, omdat er geen toets op afnamemoment E5 is ontwikkeld in LVS 2.0.

In het normeringsonderzoek M5 zijn 202 items voorgelegd aan 1767 leerlingen medio groep 5, verdeeld over de 7 boekjes zoals aangegeven in de laatste kolom van figuur 4.3. Elk boekje bestond uit 75 of 76 opgaven verdeeld over 3 taken. De 25 opgaven in de LVS 2.0 taak werden door alle 1767 leerlingen gemaakt. De overige opgaven kwamen in twee boekjes voor en werden gemiddeld door 505 leerlingen gemaakt. Bij alle opgaven werd voldaan aan de eis dat deze minimaal bij 150 leerlingen moesten zijn afgenomen.

In het normeringsonderzoek E5 zijn 127 items voorgelegd aan 1785 leerlingen eind groep 5, verdeeld over de 5 boekjes zoals aangegeven in de laatste kolom van figuur 4.4. Elk boekje bestond uit 50 tot 52 opgaven verdeeld over 2 taken. De opgaven kwamen in twee boekjes voor en werden gemiddeld door 714 leerlingen gemaakt. Bij alle opgaven werd voldaan aan de eis dat deze minimaal bij 150 leerlingen moesten 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 E4M5, M5 en E5.

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

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

als een ‘afdoende statistiek’ (*sufficient statistic*) voor de vaardigheid $\theta$. 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, $p(+\mid s)$, vergeleken worden met de feitelijk geobserveerde proportie goede antwoorden, $prop(+\mid s)$. Via de basisvergelijking van OPLM kunnen we eenvoudig de conditionele kans op het goed beantwoorden van de items afleiden en daarmee kunnen we $p(+\mid s)$ evalueren, $prop(+\mid s)$ volgt uit de data. Discrepanties tussen $p(+\mid s)$ en $prop(+\mid s)$ duiden op schendingen van het model. Deze discrepanties vormen de basis voor de diverse statistische toetsen in OPLM. De toetsingsgrootheid voor de veronderstelde discriminatie-indices is gegeven door

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

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

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

Deze zogeheten S-toets heeft een $\chi^2$ 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:

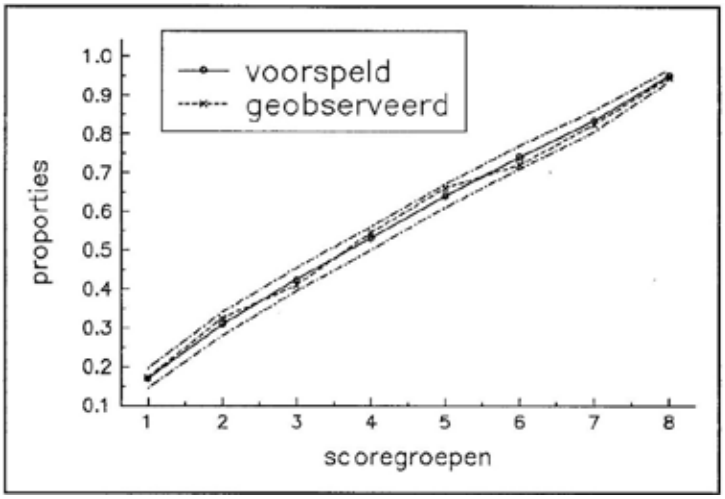
1 Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
2 Vervolgens schatten we de itemparameters met behulp van de CML-methode.
3 Met behulp van de M-toetsen controleren we of de discriminatie-indices 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 bewerkstellingen 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 stippellijn verbonden. De volle lijn daarentegen verbindt de proporties die op grond van de parameterschattingen voorspeld kunnen worden. De twee buitenste lijnen geven het 95%-betrouwbaarheidsinterval aan.
De breedte van dit interval is in belangrijke mate afhankelijk van het aantal leerlingen dat de opgave heeft beantwoord. Uit de figuur blijkt heel duidelijk dat de geobserveerde proporties, zoals bedoeld, binnen het 95%- betrouwbaarheidsinterval van de (geschatte) voorspelde proporties liggen, en dit komt in grote lijnen overeen met een niet-significante S-toetsingsgrootheid (Verhelst, et al., 1994).

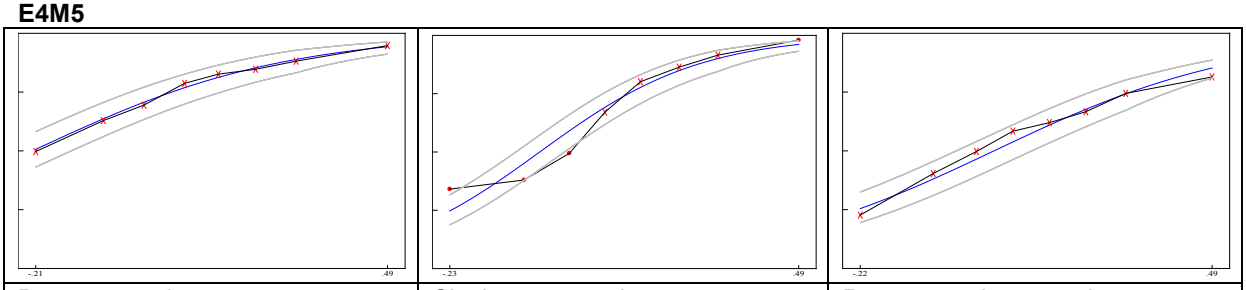
*Figuur 4.5 Grafische voorstelling van een Si –toets*




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 E4M5, M5 en E5 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.

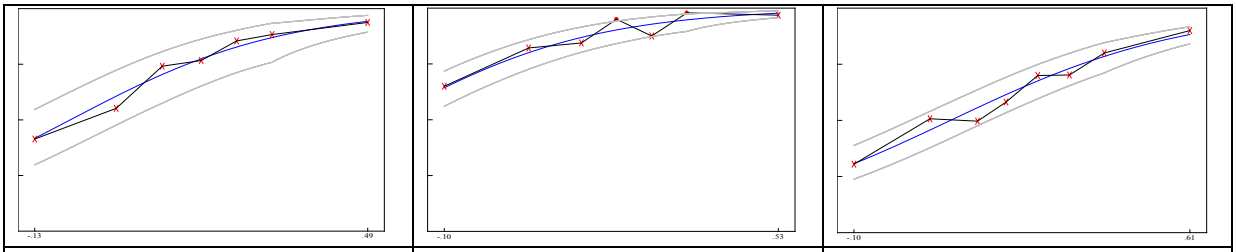
*Figuur 4.6 Voorbeelden van S-toetsen voor de Toets LVS 3.0 Begrijpend lezen E4M5, M5 en E5 met de best passende, de slechtst passende en een qua passing representatieve opgave*

**E4M5**



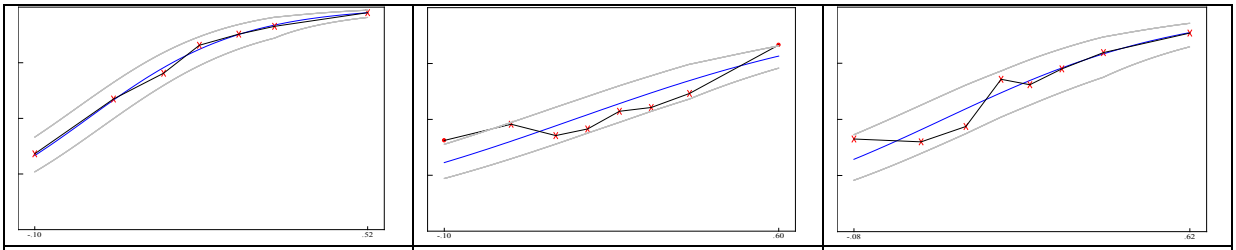
| Best passend | Slechtst passend | Representatieve passing |
|---|---|---|

**M5**



| Best passend | Slechtst passend | Representatieve passing |
|---|---|---|

**E5**



| Best passend | Slechtst passend | Representatieve passing |
|---|---|---|

In feite kan men bij de kalibratie beter varen op deze grafische weergaven dan op toetsingsresultaten in termen van exacte getallen en de significantie daarvan. 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 E4M5, M5 en E5. 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 E4M5, M5 en E5*

| | 0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| E4M5 | 3 | 3 | 2 | 4 | 3 | 5 | 5 | 5 | 5 | 3 | 9 | 3 |
| | **0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.** | | | | | | | | | | |
| M5 | 0 | 2 | 2 | 6 | 3 | 4 | 5 | 6 | 6 | 8 | 5 | 3 |
| | **0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.** | | | | | | | | | | |
| E5 | 1 | 5 | 3 | 5 | 1 | 5 | 7 | 4 | 3 | 2 | 4 | 10 |

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 E4M5, M5 en E5 geldt dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootheid is significant voor de toetsen E4M5 en E5. Aan dit laatste moet bij steekproeven met dergelijke omvang niet te veel waarde worden gehecht.

*Tabel 4.2 R1c-waarden voor E4M5, M5 en E5*

| **Toetsversie** | **R1c** | **df** | **p** |
|---|---|---|---|
| E4M5 | 501.086 | 383 | <.01 |
| M5 | 497.587 | 548 | .94 |
| E5 | 453.010 | 372 | <.002 |

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** |
| E4M5 | .04 - .11 | .06 |
| M5 | .06 - .20 | .10 |
| E5 | .05 - .14 | .09 |

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen LVS 3.0 Begrijpend lezen E4M5, M5 en E5 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-volgsysteemtoetsen 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 normerings-onderzoek, 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 dekkingsgraad van de LVS toetsen Begrijpend lezen 2$^{e}$ generatie is bijzonder hoog: de toetsen worden door 85% tot 90% van de scholen toegepast. Dit betekent dat representatieve steekproeven voor de normering van de toetsen Begrijpend lezen 3.0 in groep 5 normeringsresultaten zullen opleveren die niet of nauwelijks zullen afwijken van wat men voor de totale populatie van scholen zou mogen verwachten. In eerdere normeringsonderzoeken voor groep 3 en 4 is dat voor meerdere jaargangen nagegaan. Niet alleen de gemiddelde score op de Cito Eindtoets Basisonderwijs, maar ook specifiek de score voor begrijpend lezen bleek voor de scholen in deze normeringsonderzoeken niet af te wijken van het populatiegemiddelde van Begrijpend lezen voor de Eindtoets.

Scholen werd gevraagd om op drie opeenvolgende momenten deel te nemen. Van de deelnemende scholen op afnamemoment E4 zijn enkele scholen benaderd als herhalingsscholen voor het normerings-onderzoek op afnamemoment M5. Van de deelnemende scholen op afnamemoment M5 zijn vervolgens enkele scholen benaderd als herhalingsscholen voor het normeringsonderzoek E5. De normeringsgroep van zowel afnamemoment M5 als afnamemoment E5 bestond dus deels uit herhalingsscholen.

Voor het totaal aantal scholen in Nederland (7168 scholen) is een indeling gemaakt naar LOVS strata[5] (0 t/m 10 procent, 11 t/m 25 procent, 26 t/m 40 procent en 41 procent of meer gewichtsleerlingen) 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 M5 zijn uiteindelijk 115 herhalingsscholen en 462 extra scholen aangeschreven. Omdat na de eerste aanschrijvingsronde 56% van de herhalingsscholen uit normeringsonderzoek E4 en slechts 3% van de extra scholen uit de steekproef bereid bleek deel te nemen aan het normeringsonderzoek, zijn in een tweede wervingsronde 1506 extra scholen aangeschreven. Uiteindelijk resulteerde dit in een deelnamebereidheid van zo'n 2% van de overige aangeschreven scholen. In totaal meldden zich 108 scholen aan voor het normeringsonderzoek M5. Uiteindelijk was het aantal scholen dat daadwerkelijk gegevens aanleverde gelijk aan 76 en het aantal leerlingen aan 1767. Van de 1767 leerlingen die hebben deelgenomen aan het normeringsonderzoek konden 1607 leerlingen van 71 verschillende basisscholen worden meegenomen in de data-analyses.

Voor het normeringsonderzoek E5 zijn uiteindelijk in totaal 108 herhalingsscholen aangeschreven en 89 extra scholen. Omdat na de eerste aanschrijvingsronde 72% van de herhalingsscholen uit het normeringsonderzoek M5 ook bereid bleek deel te nemen aan het normeringsonderzoek E5 en slechts 4% van de scholen uit de aanvullende steekproef bereid bleek deel te nemen aan het normeringsonderzoek, zijn in een tweede wervingsronde 844 extra scholen aangeschreven. Uiteindelijk resulteerde dit in een deelnamebereidheid van zo'n 3% van de overige aangeschreven scholen. In totaal meldden zich 104 scholen aan voor het normeringsonderzoek. Uiteindelijk was het aantal scholen dat daadwerkelijk gegevens aanleverde gelijk aan 74 en het aantal leerlingen gelijk aan 1785. Van de 1785 leerlingen die hebben deelgenomen aan het normeringsonderzoek konden 1724 leerlingen van 71 verschillende basisscholen worden meegenomen in de data-analyses.

Voor het bepalen van de normering voor afnamemoment M5 werden de gegevens aangevuld met gegevens uit Cito dataretour. Voor afnamemoment E5 was dit niet mogelijk omdat de tweede generatie toetsen geen toets voor afnamemoment E5 kent. Hieronder beschrijven we de procedure zoals die voor het normeringsonderzoek M5 is gehanteerd. Voor E5 bestond het normeringsbestand uitsluitend uit data die verzameld werden in het *embedded field* normeringsonderzoek.
Vanzelfsprekend werden de data die via Cito dataretour binnenkwamen opgeschoond voordat ze gebruikt werden. Uit de bestanden werden 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, werden behouden.

Daarnaast werden 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 werd naar gestreefd om de uiteindelijke normeringssteekproef voor ongeveer

[5] De term stratum wordt hier gedefinieerd zoals gebruikelijk in periodieke peilingsonderzoeken, namelijk als een indicatie van de aard van de schoolpopulatie.

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 achterstandsleerlingen. Scholen zijn ingedeeld naar het percentage achterstandsleerlingen volgens een indeling in vier typen: (1) percentage achterstandsleerlingen [0, .10), (2) percentage achterstandsleerlingen [.10, .25), (3) percentage achterstandsleerlingen [.25, .40) en (4) percentage achterstandsleerlingen [.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 × 2 × 4 × 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 $N_{\mathrm{S}}$, het gewenste aantal leerlingen $N_{\mathrm{T}}$, en het beschikbare aantal leerlingen in Cito dataretour $N_{\mathrm{D}}$:

$$w_{ijk} = \frac{(n_{\mathrm{T},\,ijk} - n_{\mathrm{S},\,ijk}) \div (N_{\mathrm{T}} - N_{\mathrm{S}})}{n_{\mathrm{D},\,ijk} \div N_{\mathrm{D}}} = \frac{N_{\mathrm{D}}(n_{\mathrm{T},\,ijk} - n_{\mathrm{S},\,ijk})}{n_{\mathrm{D},\,ijk}(N_{\mathrm{T}} - N_{\mathrm{S}})},$$

waarbij vereist is dat $n_{\mathrm{S},\,ijk} \leq n_{\mathrm{T},\,ijk}$. 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 $n_{\mathrm{S},\,ijk} > n_{\mathrm{T},\,ijk}$ 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 $N_{\mathrm{T}}$ en de via de landelijke gegevens van DUO en/of CBS te bepalen $n_{\mathrm{T},\,ijk}$ 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 $n_{\mathrm{S},\,ijk} \leq n_{\mathrm{T},\,ijk}$. 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 $N_{\mathrm{S}}$ dichterbij het gewenste aantal leerlingen $N_{\mathrm{T}}$ komt te liggen. De oplossing is om $N_{\mathrm{T}}$ 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 $N_{\mathrm{S}} \geq N_{\mathrm{T}}$. 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 normeringssteekproef 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.

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 normeringssteekproef 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 normeringssteekproef de kans $w_q$ bepaald dat een leerling met sekse $q$ in een representatieve normeringssteekproef zit:

$$w_q = \frac{n_{\mathrm{T},\,q} \div N_{\mathrm{T}}}{n_{\mathrm{S},\,q} \div N_{\mathrm{S}}} = \frac{n_{\mathrm{T},\,q} N_{\mathrm{S}}}{N_{\mathrm{T}} n_{\mathrm{S},\,q}}.$$

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 normerings-steekproef 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 ($\mathrm{S}$) en Cito dataretour ($\mathrm{D}$) dus als volgt te werk:

*Voorbereiding data normeringsonderzoek*

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

*Toevoegen data uit Cito dataretour*

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

*Check leerlingvariabele sekse*

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

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 5 ingesteld op $N_{\mathrm{T}} = 2 \times 1607 = 3214$. Voor eind groep 5 kon geen dataretour worden toegevoegd. In tabel 4.4 is te zien in welke aantallen scholen en leerlingen het selectiealgoritme heeft geresulteerd. De conclusie is dat het voor afnamemoment medio groep 5 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 aselecte 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. In tabel 4.4 zijn ook de aantallen scholen en leerlingen vermeld voor het normeringsonderzoek voor afnamemoment E5. De aantallen betreffen uitsluitend nieuw verzamelde data en geen gegevens uit dataretour.

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

| Afnamemoment | Aantal leerlingen | | | Aantal scholen |
|---|---|---|---|---|
| | Steekproef | Dataretour | Totaal normering | Normering |
| M5 | 1607 | 1940 | 3547 | 157 |
| E5 | 1724 | -- | 1724 | 71 |

### 4.3.2 Representativiteit

Door de werkwijze die wordt gevolgd tijdens de normering is representativiteit van de normeringssteekproef van afnamemoment M5 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. Ook de representativiteit van de normeringssteekproef van afnamemoment E5 zal worden onderzocht. 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 M5*

| | | | Steekproef M5 | |
|---|---|---|---|---|
| **Variabele** | **Categorie** | **Populatie (%)** | **N** | **%** |
| Regio | Noord | 10.1 | 348 | 9.8 |
| | Oost | 22.5 | 803 | 22.6 |
| | West | 47.6 | 1728 | 48.7 |
| | Zuid | 19.8 | 668 | 18.8 |
| Urbanisatiegraad | Platteland | 55.3 | 1967 | 56.5 |
| | Stad | 44.7 | 1580 | 44.5 |
| Schooltype | [0, .10] | 65.0 | 2357 | 66.5 |
| | [.10, .25] | 23.2 | 822 | 23.2 |
| | [.25, .40] | 6.9 | 214 | 6.0 |
| | [.40, 1] | 4.9 | 154 | 4.3 |
| Sekse | Jongen | 50.4 | 1668 | 50.2 |
| | Meisje | 49.6 | 1658 | 49.9 |

*Tabel 4.6 Representativiteitsanalyse LVS 3.0 Begrijpend lezen E5*

| | | | Steekproef E5 | |
|---|---|---|---|---|
| **Variabele** | **Categorie** | **Populatie (%)** | **N** | **%** |
| Regio | Noord | 10.1 | 156 | 9.0 |
| | Oost | 22.5 | 272 | 15.8 |
| | West | 47.6 | 904 | 52.4 |
| | Zuid | 19.8 | 392 | 22.7 |
| Urbanisatiegraad | Platteland | 55.3 | 894 | 51.9 |
| | Stad | 44.7 | 830 | 48.1 |
| Schooltype | [0, .10] | 65.0 | 1082 | 62.8 |
| | [.10, .25] | 23.2 | 404 | 23.4 |
| | [.25, .40] | 6.9 | 137 | 8.0 |
| | [.40, 1] | 4.9 | 101 | 5.9 |
| Sekse | Jongen | 50.5 | 777 | 50.4 |
| | Meisje | 49.5 | 766 | 49.6 |

Te zien valt dat voor het afnamemoment M5 de steekproefverdeling weinig afwijkt van de populatie-verdeling. Voor afnamemoment E5 zijn de afwijkingen groter. Voor beide afnamemomenten zijn de afwijkingen statistisch getoetst. De gegevens van deze statistische toetsen worden getoond in tabel 4.7 en 4.8. De chi kwadraat-waarden zijn laag voor afnamemoment M5 en E5, met uitzondering van de waarde voor regio op afnamemoment E5. Voor afnamemoment M5 is geen enkele waarde significant, maar voor afnamemoment E5 zijn op een na alle waarden significant. Bij grotere steekproeven zegt significantie echter niet zoveel. Het is beter om de effectgrootte $\phi = \sqrt{\frac{\chi^2}{N}}$ als uitgangspunt te nemen. We zien dat de effectgroottes voor afnamemoment M5 ver onder .10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). Met .045 is de effectgrootte het grootst voor de variabele schooltype. Voor afnamemoment E5 zijn de effectgroottes voor urbanisatiegraad, schooltype en sekse eveneens klein. Voor regio op afnamemoment E5 is de effectgrootte aanzienlijk groter, namelijk .174. De conclusie luidt dat de normeringssteekproef voor afnamemoment M5 een zeer goede afspiegeling vormt van de populatie. Voor afnamemoment E5 geldt dit eveneens, uitgezonderd voor de variabele regio. Bij de bepaling van de normering is voor deze variabele gewogen. De wegingsfactoren zijn gelijk aan: noord 1.112, oost 1.427, west 0.908 en zuid 0.870. Deze waarden voldoen aan de criteria uit het COTAN-beoordelingssysteem (Evers et al., 2010)", waarin wordt aangeven dat oververtegenwoordiging geen probleem is en bij ondervertegenwoordiging een factor 2 acceptabel is.

*Tabel 4.7 Statistische toetsingsrepresentativiteitsanalyse LVS 3.0 Begrijpend lezen M5*

| | Steekproef M5 | | | | |
|---|---|---|---|---|---|
| **Variabele** | $\chi^2$ | **df** | **N** | **p** | $\phi$ |
| Regio | 2.762 | 3 | 3547 | .430 | .028 |
| Urbanisatiegraad | .033 | 1 | 3547 | .856 | .003 |
| Schooltype | 7.164 | 3 | 3547 | .067 | .045 |
| Sekse | .098 | 1 | 3326 | .754 | .005 |

*Tabel 4.8 Statistische toetsingsrepresentativiteitsanalyse LVS 3.0 Begrijpend lezen E5*

| | Steekproef E5 | | | | |
|---|---|---|---|---|---|
| **Variabele** | $\chi^2$ | **df** | **N** | **p** | $\varphi$ |
| Regio | 52.364 | 3 | 1724 | .000 | .174 |
| Urbanisatiegraad | 8.272 | 1 | 1724 | .004 | .069 |
| Schooltype | 7.402 | 3 | 1724 | .060 | .066 |
| Sekse | .007 | 1 | 1724 | .004 | .069 |

### 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 voor afnamemoment M5 de combinatie van het *embedded field* normeringsonderzoek en Cito dataretour (voor aantallen, zie tabel 4.4). Om de scores die leerlingen behalen te kunnen vergelijken over de tijd worden vaardigheidsscores gebruikt. Uit de ruwe scores van de leerlingen 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. Voor afnamemoment E5 vindt hierbij een weging plaats. In paragraaf 3.3 is de verdeling van “plausible values” voor de afnamemomenten M5 en E5 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 geeft de normgegevens op leerlingniveau.

Tabel 4.9 Normtabel op leerlingniveau voor LVS 3.0 Begrijpend lezen groep 5.

| **Afname moment** | ***M*** | ***SD*** | ***K*** | ***S*** | ***P*10** | ***P*20** | ***P*25** | ***P*40** | ***P*50** | ***P*60** | ***P*75** | ***P*80** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M5 | 155.02 | 26.20 | -.18 | -.00 | 120.34 | 132.33 | 137.06 | 148.62 | 155.46 | 162.01 | 173.00 | 177.36 |
| E5 | 159.51 | 24.57 | -.21 | .06 | 127.32 | 138.21 | 142.63 | 153.13 | 159.33 | 165.78 | 176.14 | 180.21 |

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 M5 en E5. 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 M5*

| Replicatie | Aantal scholen | Aantal leerlingen | Gemiddelde | SD School | SD Leerling | ICC |
|---|---|---|---|---|---|---|
| 1 | 159 | 3735 | 123.677 | 8.737 | 22.336 | .133 |
| 2 | 159 | 3434 | 123.368 | 9.407 | 22.033 | .154 |
| 3 | 159 | 3851 | 124.360 | 8.612 | 22.628 | .127 |
| 4 | 159 | 3760 | 123.619 | 10.400 | 22.959 | .170 |
| 5 | 159 | 3545 | 124.235 | 9.233 | 22.515 | .144 |
| … | | | | | | |
| 15 | 159 | 3739 | 123.192 | 9.550 | 22.730 | .150 |
| 16 | 159 | 3398 | 124.387 | 10.205 | 22.352 | .172 |
| 17 | 159 | 3471 | 123.342 | 10.203 | 22.174 | .175 |
| 18 | 159 | 3291 | 124.589 | 9.186 | 21.376 | .156 |
| 19 | 159 | 3319 | 123.572 | 10.325 | 22.256 | .177 |
| 20 | 159 | 3361 | 123.347 | 8.306 | 22.070 | .124 |
| Eindresultaat | | | 123.892 | 9.575 | 22.327 | 22.327 |

*Tabel 4.11 Uitkomsten multilevelanalyse voor LVS 3.0 Begrijpend lezen afnamemoment E5*

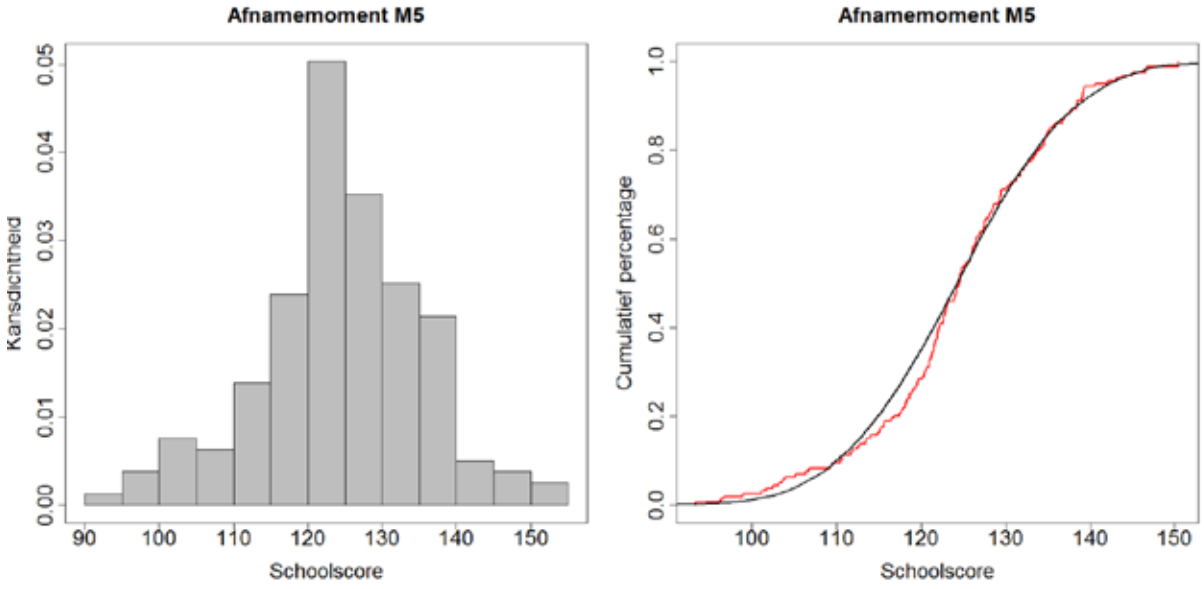
| Replicatie | Aantal scholen | Aantal leerlingen | Gemiddelde | SD School | SD Leerling | ICC |
|---|---|---|---|---|---|---|
| 1 | 71 | 1930 | 128.458 | 8.095 | 22.177 | .118 |
| 2 | 71 | 1718 | 127.087 | 7.499 | 21.697 | .107 |
| 3 | 71 | 1776 | 129.345 | 9.510 | 21.615 | .162 |
| 4 | 71 | 1787 | 128.876 | 7.012 | 21.476 | .096 |
| 5 | 71 | 1614 | 127.338 | 8.981 | 20.751 | .158 |
| … | | | | | | |
| 15 | 71 | 1572 | 128.518 | 10.813 | 21.015 | .209 |
| 16 | 71 | 1704 | 128.673 | 8.064 | 20.733 | .131 |
| 17 | 71 | 1639 | 125.919 | 8.772 | 20.558 | .154 |
| 18 | 71 | 1678 | 129.510 | 6.813 | 21.907 | .088 |
| 19 | 71 | 1832 | 129.164 | 9.325 | 21.246 | .162 |
| 20 | 71 | 1731 | 128.269 | 10.180 | 21.277 | .186 |
| Eindresultaat | | 128.054 | 8.422 | 22.327 | 21.412 | .136 |

Figuur 4.7 en 4.8 laten de verdeling van schoolgemiddelden M5 en E5 zien. Het is lastig te bepalen of de schoolgemiddelden een normale verdeling volgen met een scholenaantal van 159 en 71. 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 vaardigheids-verdeling 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 5 terwijl de gemiddelden en standaarddeviaties in tabel 4.12 op de (getransformeerde) overkoepelende vaardigheidsschaal worden weergegeven.

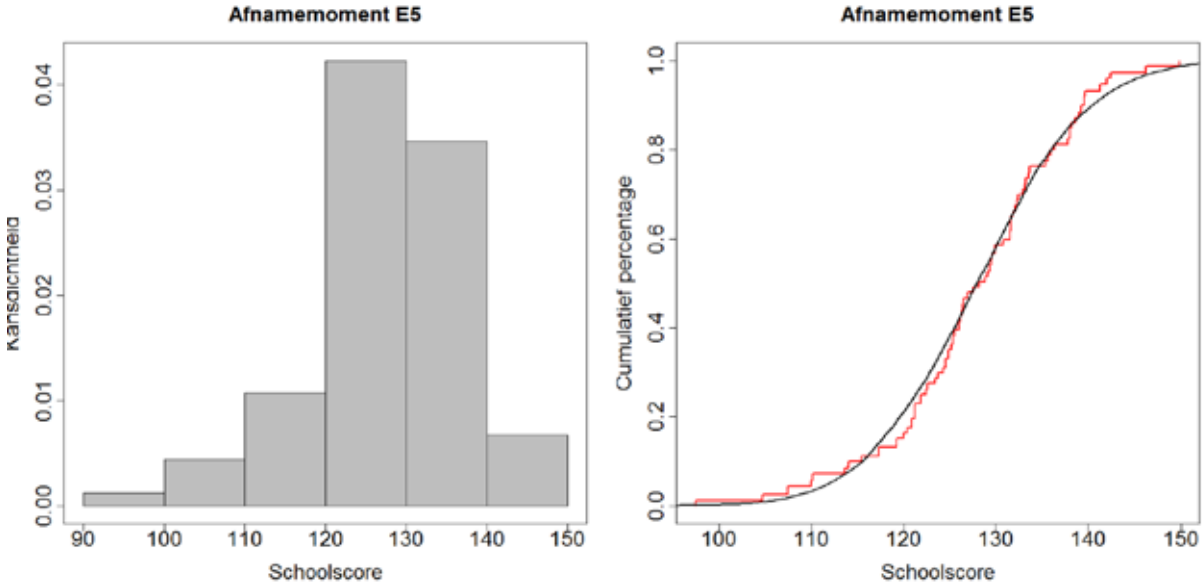
*Tabel 4.12 Normtabel op schoolniveau voor LVS 3.0 Begrijpend lezen groep 5*

| Afname Moment | *M* | *SD* | *P*10 | *P*20 | *P*25 | *P*40 | *P*50 | *P*60 | *P*75 | *P*80 |
|---|---|---|---|---|---|---|---|---|---|---|
| M5 | 154.34 | 10.95 | 140.31 | 145.13 | 146.96 | 151.57 | 154.34 | 157.12 | 161.73 | 163.56 |
| E5 | 159.10 | 9.63 | 146.76 | 151.00 | 152.61 | 156.66 | 159.10 | 161.54 | 165.60 | 167.21 |

*Figuur 4.7 Verdeling van de schoolgemiddelden voor LVS 3.0 Begrijpend lezen M5*




*Figuur 4.8 Verdeling van de schoolgemiddelden voor LVS 3.0 Begrijpend lezen E5*



### 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 5 een geldigheid aanhouden tot en met 2024.
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 E4M5, M5 en E5 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 gebruik gemaakt van het programma OPLAT (Verhelst, Glas & Verstralen, 1995).

In OPLAT wordt een door Verhelst et al. (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 τ(θ). 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(τ). Tussen θ en τ(θ) bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de toetsscore τ(θ) moet behalen (dat is alleen zo als de toets oneindig lang wordt).
De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor 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|τ(θ)), 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.

$$Var(t) = E[Var(t \mid \tau(\theta))] \tag{5.1}$$

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

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

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 totaal aantal opgaven dat deel uitmaakt van de toetsen E4M5, M5 en E5. 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: de waarden variëren tussen .88 en .90.

*Tabel 5.1 Beschrijvende gegevens bij de toetsen Begrijpend lezen E4M5, M5 en E5*

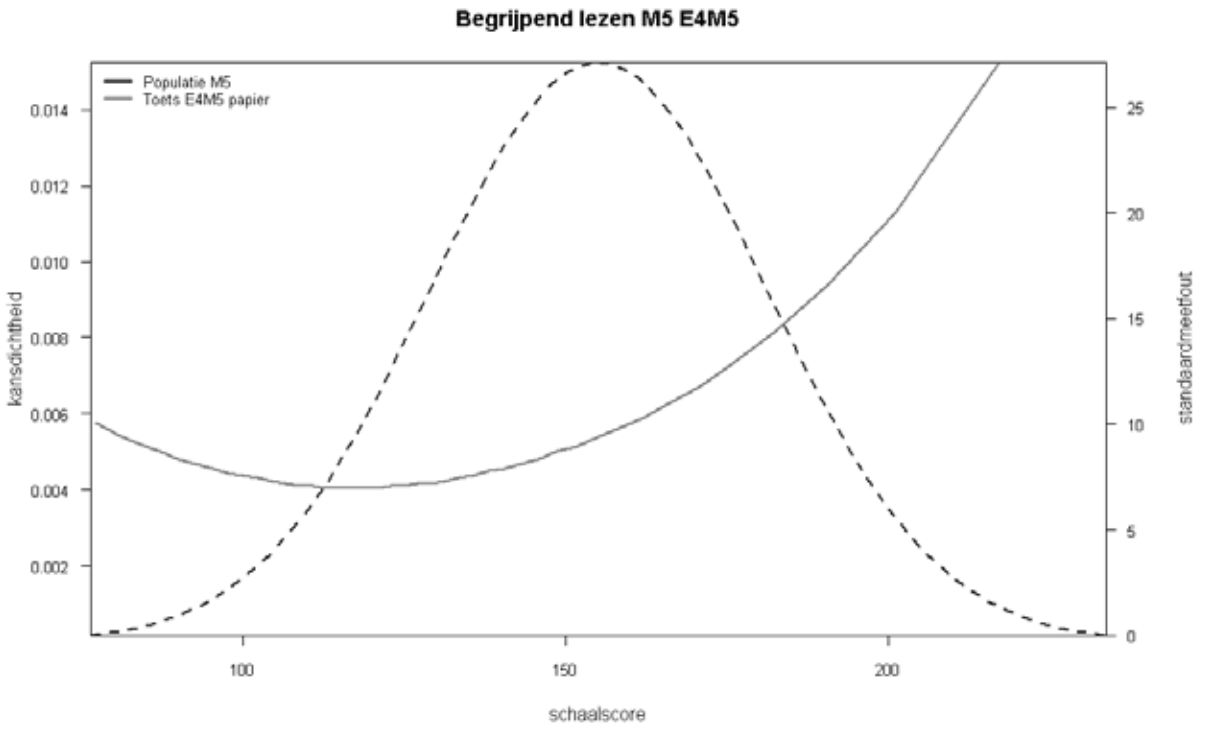
| Toets | Maximum score | Gemiddelde | Standaard-meetfout | MAcc | Test-hertest (simulatie) |
|---|---|---|---|---|---|
| E4M5 | 50 | 37.9 | 2.715 | .90 | .90 |
| M5 | 50 | 35.5 | 2.885 | .89 | .89 |
| E5 | 50 | 34.4 | 2.978 | .88 | .88 |

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 aselect 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 geen 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 toetsen Begrijpen lezen.

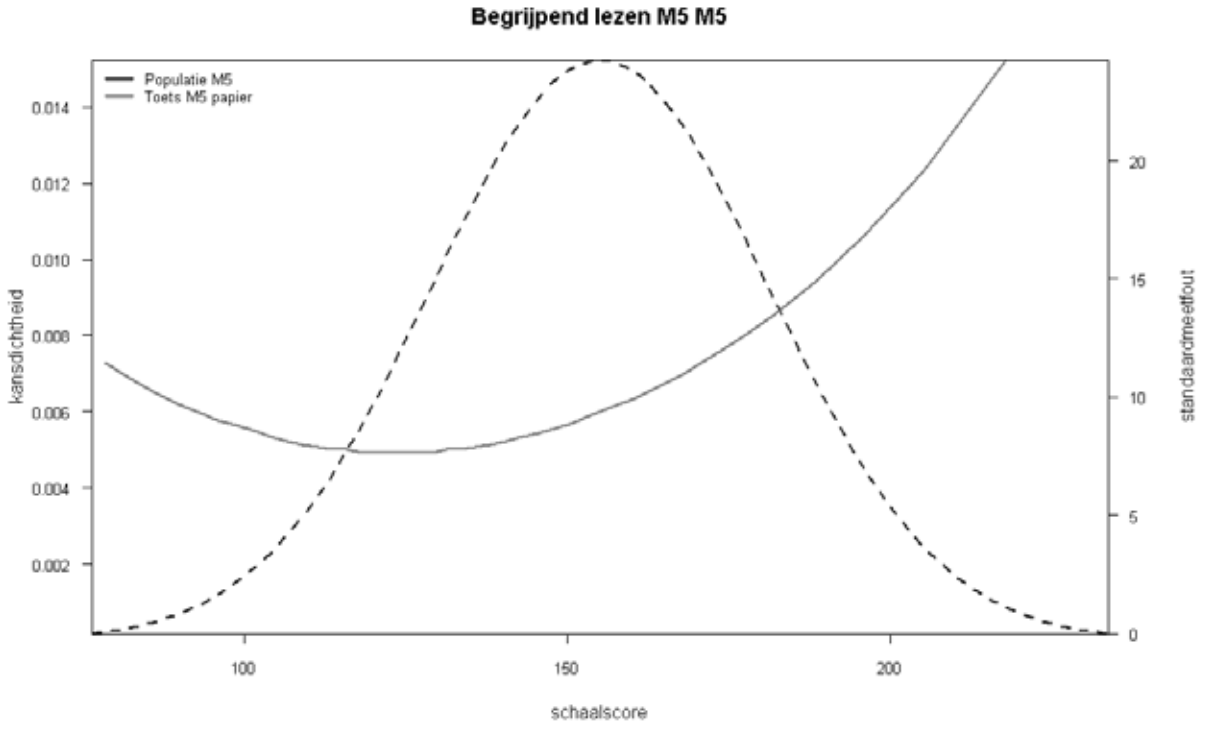
## 5.2 Nauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meet-nauwkeurigheid van de toetsen Begrijpend lezen E4M5, M5 en E5 en geven geen beeld van de lokale meetnauwkeurigheid. Figuren 5.1 tot en met 5.3 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 afname-moment opgenomen. Deze laatste 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.

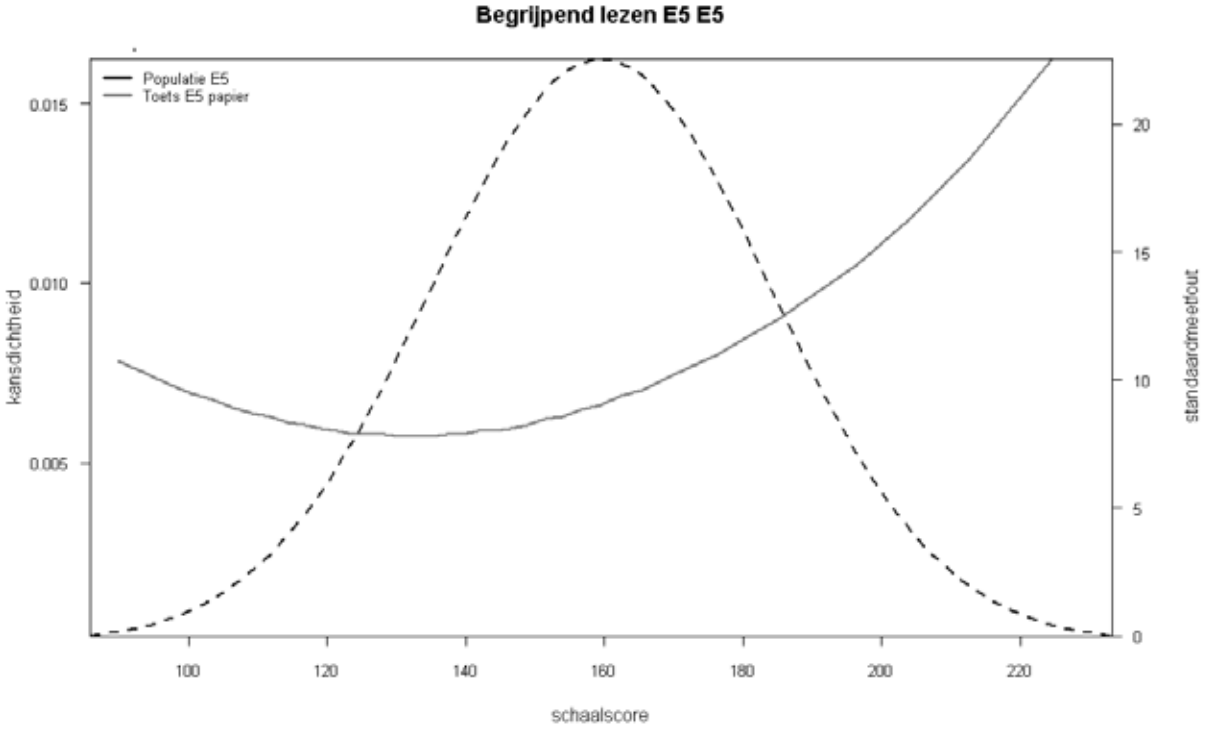
*Figuur 5.1 Grootte van de meetfouten voor de toets E4M5 en de kansdichtheidsfunctie voor de M5-populatie*




*Figuur 5.2 Grootte van de meetfouten voor de toets M5 en de kansdichtheidsfunctie voor de M5-populatie*



*Figuur 5.3 Grootte van de meetfouten voor de toets E5 en de kansdichtheidsfunctie voor de E5-populatie*




**Betrouwbaarheidstabellen**

De betekenis van de (lokale) meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit betrouwbaarheidstabellen. De tabellen 5.5a tot en met 5.5c laten voor de toetsen E4M5, M5 en E5, respectievelijk afnamemomenten medio groep 5 en einde groep 5, zien hoe vaak de werkelijke vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat tabel 5.5a zien dat 87,9 procent van de leerlingen die halverwege groep 5 op basis van de E4M5-toets in scoregroep V geclassificeerd wordt ook met hun werkelijke vaardigheidsscore in deze scoregroep geclassificeerd wordt. De kans dat een V-leerling terecht als V-leerling wordt bestempeld is, met andere woorden, ongeveer 88 procent. Verder laat de linkerkant van Tabel 5.5a zien dat 12 procent van de leerlingen in scoregroep V een vaardigheidsscore heeft die in werkelijkheid in scoregroep IV valt. De overige getallen in tabellen 5.5a tot en met 5.5c zijn op dezelfde wijze te interpreteren.

*Tabel 5.5a Betrouwbaarheidstabel Toets E4M5 voor afnamemoment medio 5*

| | **Scoregroep waarin de ware score valt** | | | | | | **Scoregroep waarin de ware score valt** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Score-groepen V t/m I** | **V** | **IV** | **III** | **II** | **I** | **Score-groepen E t/m A** | **E** | **D** | **C** | **B** | **A** |
| V | 87,9 | 12,0 | 0,1 | 0,0 | 0,0 | E | 85,0 | 15,0 | 0,1 | 0,0 | 0,0 |
| IV | 17,5 | 62,3 | 19,2 | 0,9 | 0,0 | D | 14,3 | 67,0 | 18,6 | 0,1 | 0,0 |
| III | 0,8 | 23,4 | 51,6 | 22,5 | 1,6 | C | 0,2 | 16,9 | 64,0 | 18,4 | 0,5 |
| II | 0,1 | 3,5 | 24,7 | 46,8 | 25,0 | B | 0,0 | 0,8 | 23,2 | 54,4 | 21,7 |
| I | 0,0 | 0,5 | 4,5 | 19,6 | 75,5 | A | 0,0 | 0,1 | 2,5 | 19,0 | 78,5 |

*Tabel 5.5b Betrouwbaarheidstabel Toets M5 voor afnamemoment medio 5*

| | Scoregroep waarin de ware score valt | | | | | | Scoregroep waarin de ware score valt | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Score-groepen V t/m I** | **V** | **IV** | **III** | **II** | **I** | **Score-groepen E t/m A** | **E** | **D** | **C** | **B** | **A** |
| V | 87,3 | 12,6 | 0,2 | 0,0 | 0,0 | E | 83,7 | 16,1 | 0,2 | 0,0 | 0,0 |
| IV | 17,7 | 61,9 | 19,4 | 0,9 | 0,0 | D | 14,9 | 65,7 | 19,2 | 0,1 | 0,0 |
| III | 0,8 | 23,2 | 52,2 | 22,4 | 1,5 | C | 0,2 | 16,8 | 64,0 | 18,4 | 0,5 |
| II | 0,0 | 3,0 | 24,4 | 48,2 | 24,3 | B | 0,0 | 0,6 | 22,7 | 55,5 | 21,2 |
| I | 0,0 | 0,3 | 3,6 | 19,2 | 76,9 | A | 0,0 | 0,0 | 1,9 | 18,3 | 79,8 |

*Tabel 5.5c Betrouwbaarheidstabel Toets E5 voor afnamemoment einde 5*

| | Scoregroep waarin de ware score valt | | | | | | Scoregroep waarin de ware score valt | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Score-groepen V t/m I** | **V** | **IV** | **III** | **II** | **I** | **Score-groepen E t/m A** | **E** | **D** | **C** | **B** | **A** |
| V | 86,1 | 13,6 | 0,3 | 0,0 | 0,0 | E | 82,3 | 17,2 | 0,4 | 0,0 | 0,0 |
| IV | 18,4 | 59,2 | 20,6 | 1,9 | 0,0 | D | 18,0 | 60,6 | 21,1 | 0,3 | 0,0 |
| III | 1,1 | 25,3 | 47,6 | 24,6 | 1,4 | C | 0,4 | 16,9 | 62,8 | 19,4 | 0,4 |
| II | 0,1 | 3,7 | 22,6 | 50,2 | 23,4 | B | 0,0 | 0,7 | 22,0 | 57,0 | 20,3 |
| I | 0,0 | 0,2 | 2,7 | 18,5 | 78,6 | A | 0,0 | 0,0 | 1,6 | 18,4 | 80,1 |

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 5 en einde groep 5 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, 97,6 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 68 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 E4M5, en dan met name bij de hoogste scoregroep. Dat is conform verwachting, aangezien deze toets – die wat makkelijker is dan de toets M5 – expliciet bedoeld is voor de minst vaardige leerlingen halverwege groep 5. 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 5 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 E4M5, M5 en E5 op afnamemomenten groep 5*

| | **Toets E4M5, afnamemoment M5** | | **Toets M5, afnamemoment M5** | | **Toets E5, afnamemoment E5** | |
|---|---|---|---|---|---|---|
| | **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 | 64.8 | 67.9 | 65.3 | 68.2 | 64.4 | 67.3 |
| Accuracy plus/minus 1 niveau | 97.6 | 99.0 | 97.9 | 99.1 | 97.7 | 99.1 |

Verdere gedetailleerde informatie over de meetnauwkeurigheid van de toets is te vinden in de handleiding van het toetspakket Begrijpend lezen groep 5 (Cito, 2015). In de schaalscoretabellen van bijlage 2 in de handleiding is een kolom opgenomen waarin het score-interval vermeld is. In deze kolom staat voor iedere ruwe score op elke toets het 67-procents-betrouwbaarheidsinterval voor de bijbehorende vaardigheidsschatting.

# 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 5, 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 aan te leggen bij de toetsconstructie. 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 de orde werd gesteld in hoofdstuk 2. In hoofdstuk 3 is deze domeinbeschrijving vervolgens nader 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 5.
De toepassing van bovengenoemde inhoudelijke criteria had tot doel evenwichtige en representatieve toetsen te construeren. Uit de manier waarop de inhoudelijke criteria zijn toegepast en uit de gerealiseerde toetsinhoud (zie hiervoor hoofdstuk 2 en 3), kunnen we afleiden dat we hierin zijn geslaagd: de inhoudsvaliditeit van de toetsen Begrijpend lezen voor E4M5, M5 en E5 is goed te noemen.

## 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 redelijk 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 toetsingsgrootheid is significant voor de toetsen E4M5 en E5, maar hier moet bij steekproeven van deze omvang niet teveel waarde aan worden gehecht.

Tenslotte 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 drie toetsen (variërend van .06 tot .10) 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. Een geslaagde kalibratie impliceert dat dezelfde onderliggende vaardigheid ten grondslag ligt aan de toetsen Begrijpend lezen E4M5, M5 en E5, hetgeen we interpreteren als een noodzakelijke voorwaarde voor begripsvaliditeit. Dat het hierbij daadwerkelijk gaat om de vaardigheid Begrijpend lezen zal blijken uit de overige analyses die we in dit hoofdstuk presenteren.

### 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 E4M5 tot en met E5. Voor de toetsen is te zien dat de p-waarden liggen tussen de .45 en .92. De p-waarden van de items liggen op twee items voor E4M5 na tussen de beoogde .40 en .90. Bij het opnemen van twee iets te gemakkelijke opgaven werden inhoudelijke overwegingen geprefereerd boven psychometrische. Daarnaast is het juist voor deze gemakkelijke toetsvariant ook niet problematisch dat er enkele iets gemakkelijker items inzitten. Het zijn immers juist de zwakkere leerlingen die deze toets zullen maken. Ook is er gezorgd voor een goede spreiding van moeilijkheid over de items. De gemiddelde moeilijkheid van de toetsen M5 en E5 is respectievelijk .71 en .69, waar het streven was een waarde te realiseren tussen .65 en .75. Daarmee zijn de toetsen niet te moeilijk en wordt voorkomen dat de leerling gefrustreerd raakt tijdens de toetsafname. De toets E4M5 is zoals beoogd iets gemakkelijker. De $R_{it}$-waarde ligt voor twee items uit de toets E4M5, zes items uit de toets M5 en twee items uit de toets E5 onder .30, maar is voor alle items groter dan .20 (in het COTAN Beoordelingssysteem de ondergrens voor de beoordeling 'voldoende'). Door de Cotan wordt een $R_{it}$-waarde groter dan .30 gekwalificeerd als goed. Met gemiddelden van .38 tot .41 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 E4M5, M5 en E5*

| | **P-waarden** | | **$R_{it}$-waarden** | | **N items** |
|---|---|---|---|---|---|
| | **Range** | **Gemiddelde** | **Range** | **Gemiddelde** | |
| E4M5 | .55-.92 | .76 | .28-.53 | .41 | 50 |
| M5 | .45-.87 | .71 | .24-.54 | .40 | 50 |
| E5 | .45-.85 | .69 | .28-.54 | .38 | 50 |

In tabel 6.2 wordt de verdelingskarakteristiek gegeven van de ruwe scores op het afnamemoment dat hoort bij de toetsen E4M5, M5 en E5. 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 M5 en E5 en iets hoger is voor de toets E4M5, 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 E4M5, M5 en E5*

| Meetmoment | Aantal opgaven | Gemiddelde | SD | Scheefheid | Kurtosis |
|---|---|---|---|---|---|
| E4M5 | 50 | 37.9 | 8.45 | -1.013 | .716 |
| M5 | 50 | 35.5 | 8.66 | -.787 | .161 |
| E5 | 50 | 34.4 | 8.54 | -.660 | -.041 |

## 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 één item in elk van de toetsen E4M5, M5 en E5 differentieel functioneren zien (bij $\alpha = .01$). Voor de toetsen Begrijpend lezen van leerjaar 5 is er dus nauwelijks sprake van itembias met betrekking tot sekse. Het aantal van drie significante toetsingen (op in totaal 150 toetsingen) ligt immers dicht in de buurt van de één à twee fouten van de eerste soort die men bij het gekozen significantieniveau mag verwachten. Daar komt nog bij dat de grafische weergaven geen bijzonderheden laten zien.

## 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 (tweede generatie) van groep 5, met de LVS 3.0-toetsen Begrijpend lezen van groep 3 en 4 en met een toets Begrijpend lezen van uitgeverij Boom. De divergente validiteit is onderzocht door de samenhangen met andere taalonderdelen en rekenen-wiskunde te onderzoeken .

*Convergente validiteit: Correlatie scores LVS tweede generatie met LVS 3.0 Begrijpend lezen*

De leerlingen van de normeringssteekproef M5 hebben zowel opgaven van de toetsen Begrijpend lezen van LVS tweede generatie 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 tweede generatie opgaven en die op basis van de LVS 3.0 opgaven. De correlatie tussen de gewogen score gebaseerd op items uit LVS tweede generatie en de gewogen score gebaseerd op items uit LVS 3.0 is hoog voor M5 ($r=.89$; $N=1767$). Aangezien de begripsvaliditeit van de toetsen Begrijpend lezen LVS tweede generatie 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. Voor de leerlingen van de normeringssteekproef E5 was een soortgelijke analyse niet mogelijk, omdat er in LVS tweede generatie geen toets voor dat afnamemoment bestaat.

*Convergente validiteit: Correlatie scores LVS 3.0 M4, E4, M5 en E5 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. In het jaar voorafgaand aan de uitgave van LVS 3.0 Begrijpend lezen groep 5 zijn de toetsen LVS 3.0 Begrijpend lezen groep 3 en groep 4 uitgegeven. Om een uitspraak te kunnen doen over de correlaties tussen leerlingen op verschillende afnamemomenten is een gedeelte van de leerlingen uit de normeringssteekproef M4 gevolgd tot op afnamemoment M5, zijn leerlingen uit normeringssteekproef E4 gevolgd tot op afnamemoment E5 en worden leerlingen uit normeringssteekproef M5 gevolgd tot op afnamemoment M6. De leerlingen die in schooljaar 2013-2014 hebben meegedaan aan een onderzoek waarin de beoogde toetsen voor M4 zijn voorgelegd zijn gevolgd tot aan moment M5. Leerlingen die in schooljaar 2013-2014 hebben meegedaan aan een onderzoek waarin de beoogde toetsen voor E4 zijn voorgelegd zijn gevolgd tot aan moment E5. Leerlingen die hebben meegedaan aan normeringsonderzoek M5 in schooljaar 2014-2015 zijn ook gevolgd op moment E5.
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 M4, E4, M5 en E5*

| | **M4** | **E4** | **M5** | **E5** |
|---|---|---|---|---|
| M4 | --- | 967 | 578 | 454 |
| E4 | .85 | --- | 1288 | 995 |
| M5 | .81 | .82 | --- | 1219 |
| E5 | .77 | .85 | .88 | --- |

Alle correlaties zijn met waarden van .77 of hoger hoog te noemen. De correlatie tussen de score op de M5-toets en de score op de E5-toets is het hoogst (.88), terwijl de correlatie tussen de scores op de M4- en de E5-toets het laagst is. Dit ligt in de lijn der verwachtingen aangezien de langste afnameperiode (1,5 jaar) tussen de afnamemomenten M4 en E5 zit. Maar ook tussen afnamemomenten M4 en M5 én E4 en E5 is deze periode met één leerjaar nog redelijk lang. Tussen afnamemoment M5 en E5 is de afnameperiode het kortst. Het interval tussen de afnamemomenten lijkt dus van invloed op de hoogte van de correlatie, hetgeen in overeenstemming is met onze verwachtingen hieromtrent.

*Convergente validiteit: Correlatie scores LVS 3.0 M5 en de Schoolvaardigheidstoets Begrijpend lezen*

Scholen zijn benaderd om op het medio-afnamemoment met groep 5 de nieuwe toets af te nemen en daarnaast de Schoolvaardigheidstoets Begrijpend lezen (Uitgeverij Boom). De Schoolvaardigheidstoets Begrijpend lezen is net als de LVS-toets Begrijpend lezen een toets met meerkeuze-opgaven over teksten. De Schoolvaardigheidstoets van groep 5 bestaat uit 25 opgaven over informatieve teksten die allemaal over hetzelfde onderwerp gaan, namelijk de otter. De Commissie Testaangelegenheden Nederland (COTAN) beoordeelde de Schoolvaardigheidstoets Begrijpend lezen op het onderdeel normen als goed. De andere onderdelen (uitgangspunten bij de testconstructie, kwaliteit van het testmateriaal, kwaliteit van de handleiding, betrouwbaarheid en begripsvaliditeit) kregen het oordeel voldoende.

Bij 89 leerlingen van 6 verschillende scholen werden op afnamemoment M5 naast twee taken met nieuwe items van de toets LVS Begrijpend lezen 3.0 M5, de 25 items uit de Schoolvaardigheidstoets Begrijpend lezen groep 5 afgenomen. De Pearson-correlatie (na correctie voor attenuatie) tussen de beide ruwe scores bleek goed: r = 0,70. We interpreteren deze samenhang als evidentie voor de soortgenootvaliditeit van de toets LVS Begrijpend lezen 3.0 M5 en (indirect) de op dezelfde schaal gekalibreerde toetsen E4M5 en E5. De correlatie is, in vergelijking met de correlaties tussen de LVS-toetsen Begrijpend lezen onderling (zie tabel 6.3) niet bijzonder hoog. Mogelijk speelt hierbij de nogal eenzijdige operationalisatie van de

Schoolvaardigheidstoets een rol. De Schoolvaardigheidstoets bestaat uitsluitend uit informatieve teksten en alle opgaven gaan over hetzelfde onderwerp (‘de otter’).

*Divergente validiteit: Correlatie scores LVS 3.0 met diverse toetsen leervorderingen*

Aan de scholen die hebben deelgenomen aan het normeringsonderzoek E5 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 leervorderingentoetsen van Cito beschikbaar.

Op basis van algemene cognitieve verschillen tussen leerlingen (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 correlatie met technisch lezen steeds minder zal worden naarmate leerlingen ouder worden.
In vergelijking met groep 3 en 4 zal de correlatie met technisch lezen daarom lager zijn. De rol van het technisch lezen gaat namelijk een steeds kleinere rol spelen, omdat leerlingen de leestechniek steeds beter gaan beheersen. 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 verwachten we een redelijke samenhang. 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. We verwachten dat de vaardigheid in begrijpend lezen in groep 5 redelijk sterk samenhangt met de woordenschat. In beide toetsen ligt sterk de nadruk op de semantiek en bij tekstbegrip speelt woordkennis een belangrijke rol. Voor luistervaardigheid, ten slotte, verwachten we een hoge samenhang. Tussen de toetsen Begrijpend lezen en Begrijpend luisteren zijn namelijk veel overeenkomsten, vooral als het gaat om vaardigheden zoals het afleiden en integreren van informatie uit een tekst. In beide toetsen komen het kunnen benoemen van de hoofdgedachte of de bedoeling van de tekst aan de orde. Alleen gebeurt dit bij de toetsen Begrijpend luisteren aan de hand van gesproken, (veelal) authentieke teksten die specifiek toegesneden zijn op het in kaart brengen van de begripsvaardigheid van gesproken materiaal.
In tabel 6.4 is de correlatie tussen de scores op de toets Begrijpend lezen E5 en de toetsen Technisch lezen (Leestempo, DMT), Spelling, Rekenen-Wiskunde, Woordenschat en Begrijpend luisteren op ditzelfde afnamemoment weergegeven.

*Tabel 6.4 Correlaties tussen Begrijpend lezen M5 en verschillende andere LVS-onderdelen*

| | **Begrijpend lezen*** | **Aantal leerlingen** |
|---|---|---|
| Cito Technisch lezen – Leestempo M5 | .51 | 156 |
| Cito Technisch lezen – DMT M5 | .40 | 589 |
| Cito Spelling M5 | .58 | 567 |
| Cito Rekenen-Wiskunde M5 | .62 | 734 |
| Cito Woordenschat M5 | .70 | 635 |
| Cito Begrijpend luisteren | .82 | 70 |

*Deze correlaties zijn gecorrigeerd voor attenuatie

Uit tabel 6.4 blijkt dat de samenhang tussen enerzijds begrijpend lezen en anderzijds begrijpend luisteren inderdaad sterk is. Het betreft de hoogste correlatie in deze tabel. Ook de correlatie met woordenschat is vrij hoog. De correlatie met technisch lezen is lager in vergelijking met groep 3 en 4. In de Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 3 (Jolink, Tomesen, Hilte, Weekers en Engelen, 2015) en de Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 4 (Jolink, Tomesen, Hilte, Weekers en Engelen, 2015) staan de correlaties vermeld tussen Begrijpend lezen en

andere toetsen in groep 3 respectievelijk groep 4. De correlatie op het E3-moment tussen begrijpend lezen en verschillende toetsen voor technisch lezen lag tussen .62 en .71 en op het E4-moment tussen .49 en .55. Ook de correlaties met spelling en rekenen-wiskunde in groep 5 zijn in overeenstemming met onze verwachtingen.

Samenvattend kan dus gesteld worden dat de correlaties van de LVS 3.0 – toetsen groep 5 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 afnamemoment M5. Ook de correlaties tussen toetsscores voor Begrijpend lezen 3.0 op verschillende afnamemomenten (M4, E4, M5 en E5) zijn hoog. Voor afnamemoment M5 zien we dat de correlatie met andere leervorderingentoetsen, die op hetzelfde moment zijn afgenomen lager is dan de correlatie voor afnamemoment M5 met LVS 2.0 M5. Alleen de correlatie tussen de Schoolvaardigheidstoets Begrijpend lezen en LVS-Begrijpend lezen 3.0 M5 hadden we iets hoger verwacht, wat mogelijk toe te schrijven is aan de verschillende manieren waarop beide toetsen zijn geoperationaliseerd.
Al met al vormen de resultaten 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 M5 en E5*

**M5**

| Leeftijdsgroep | Aantal | M | SD |
|---|---|---|---|
| 7.5 en jonger | 7 | 177.6 | 11.1 |
| 8-9 | 1321 | 157.6 | 26.3 |
| 9.5 en ouder | 168 | 139.4 | 23.1 |

**E5**

| Leeftijdsgroep | Aantal | M | SD |
|---|---|---|---|
| 8 en jonger | 8 | 187.1 | 17.1 |
| 8.5-9.5 | 1444 | 161.5 | 26.3 |
| 10 en ouder | 223 | 145.5 | 22.6 |

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 = .83 voor M5 en effectgrootte = 1.04 voor E5 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 zeven leerlingen op afnamemoment M5 en acht leerlingen op afnamemoment E5 gaat. Aan de andere kant scoort de oudste groep leerlingen het laagst (effectgrootte -.62 voor M5 en effectgrootte -.54 voor E5). 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 gemiddelde score die valt tussen de gemiddelde scores van de jongere en oudere leerlingen.

Voor de variabele geslacht is de gemiddelde score van jongens en meisjes afzonderlijk bepaald, zie tabel 6.6. De meisjes behalen een iets hogere score dan de jongens. Er is sprake van een klein effect (-.30 voor afnamemoment M5 en -.27 voor afnamemoment E5). 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 M5 en E5*

**M5**

| Geslacht | Aantal | M | SD |
|---|---|---|---|
| jongen | 688 | 151.7 | 26.0 |
| meisje | 671 | 159.7 | 26.3 |

**E5**

| Geslacht | Aantal | M | SD |
|---|---|---|---|
| jongen | 777 | 155.8 | 25.9 |
| meisje | 1006 | 162.9 | 26.2 |

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 M5 en effectgrootte .14 voor afnamemoment E5 ten opzichte van het algemene gemiddelde). Aan de andere kant scoren de leerlingen met het hoogste gewicht (1.20) het laagst (effectgrootte -.74 voor M5 en effectgrootte -.69 voor E5). 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 van wie de ouders een hogere opleiding hebben genoten (zie bijv. Kuhlemeier et al., 2014).

*Tabel 6.7 Gemiddelde score per gewicht voor de afnamemomenten M5 en E5*

**M5**

| Gewicht | Aantal | M | SD |
|---|---|---|---|
| 0.00 | 971 | 159.1 | 26.1 |
| 0.30 | 69 | 141.7 | 23.7 |
| 1.20 | 66 | 136.0 | 25.5 |

**E5**

| Geslacht | Aantal | M | SD |
|---|---|---|---|
| 0.00 | 1046 | 163.2 | 26.6 |
| 0.30 | 71 | 145.5 | 21.1 |
| 1.20 | 77 | 141.4 | 21.3 |

# 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 5 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 de andere 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 toetsen aansluit bij het referentiekader Nederlandse taal, de kerndoelen Nederlandse taal, de tussendoelen gevorderde geletterdheid en de Leerstoflijnen lezen. Deze bronnen vormden een adequate basis voor de domeinbeschrijving van de toetsen Begrijpend 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 Volgsysteem 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 toetsen uitgewerkt. Daarbij zijn de doelen voor de toetsen van groep 5 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 proeftoetsingen en de normerings-onderzoeken en de samenstelling van de definitieve toetsen. Ten slotte bevat hoofdstuk 3 een beknopte statistische beschrijving van de toetsen.

In hoofdstuk 4 rapporteerden we over de kalibratie en normering. We beschreven de opzet van en 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 normeringssteekproeven op basis van de variabelen regio, urbanisatiegraad, schooltype en sekse een goede afspiegeling vormen 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 toetsen liggen tussen .88 en .90 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 toetsen 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. Een eerste belangrijke grondslag voor de *begripsvaliditeit* is te vinden in 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 Rit-waarden zijn goed te noemen. DIF-onderzoek toont daarnaast aan dat er bij slechts één item in elk van de toetsen E4M5, M5 en E5 sprake is van itembias met betrekking tot sekse.
In hoofdstuk 6 bespraken we ook de soortgenootvaliditeit. Als belangrijke schakel in de bewijsvoering werd de hoge correlatie met de Cito LVS-toetsen Begrijpend lezen van de tweede generatie opgevoerd. Deze toetsen werden eerder door de COTAN op begripsvaliditeit positief beoordeeld. Ook de samenhangen met de toetsen Begrijpend lezen 3.0 voor groep 4 en tussen de nieuwe toetsen M5 en E5 onderling zijn hoog. De correlatie van de LVS-toetsen Begrijpend lezen 3.0 met de Schoolvaardigheidstoets Begrijpend lezen van uitgeverij Boom is, in vergelijking met de correlaties tussen de LVS-toetsen Begrijpend lezen onderling niet bijzonder hoog. Mogelijk speelt hierbij de nogal eenzijdige operationalisatie van de Schoolvaardigheidstoets een rol.
Al met al is de convergente validiteit van de nieuwe toetsen hoog te noemen. De correlaties met andere toetsen op het gebied van leervorderingen bleken lager dan de correlatie tussen de toetsen Begrijpend lezen onderling. Dit kan als bewijs van divergente validiteit worden opgevat.
Als laatste werden verschillen tussen relevante subgroepen (naar leeftijd, sekse en leerlinggewicht) gepresenteerd. De resultaten bleken aan te sluiten bij de verwachtingen die op grond van theoretische inzichten en eerder onderzoek konden worden geformuleerd en vormen daarmee extra ondersteuning voor de validiteit van de toetsen.
Op basis van deze analyses, die licht werpen op diverse aspecten van validiteit, kunnen we concluderen dat de LVS-toetsen Begrijpend lezen 3.0 voor groep 5 begripsvalide instrumenten zijn om de spellingvaardigheid te beschrijven en te volgen.

# 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.)*. Chestnutt Hill, MA: Boston College

Cito (2015). *Cito Volgsysteem primair en speciaal onderwijs. Begrijpend lezen 3.0 Groep 5*. 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.

Jolink, A., Tomesen, M., Hilte, M., Weekers, A. & Engelen, R. (2015). *Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 4*. 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.), T*heoretical 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 leesonderwijs 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.

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

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 itemparameters 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

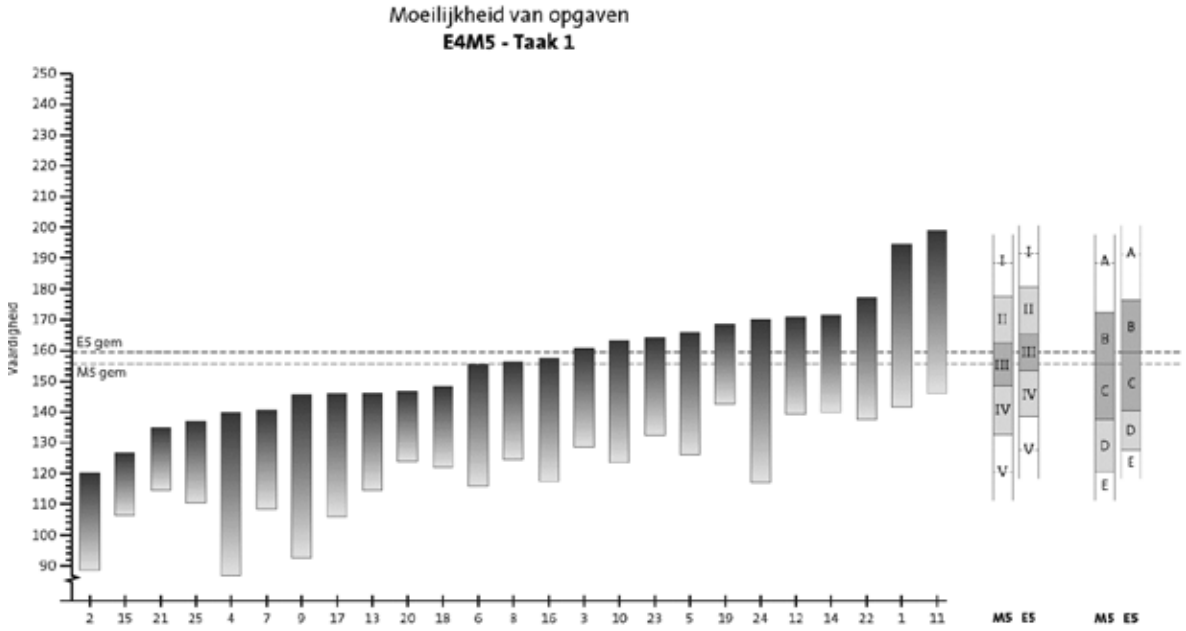
*Referentieniveaus 1F en 2F voor het lezen van zakelijke teksten*

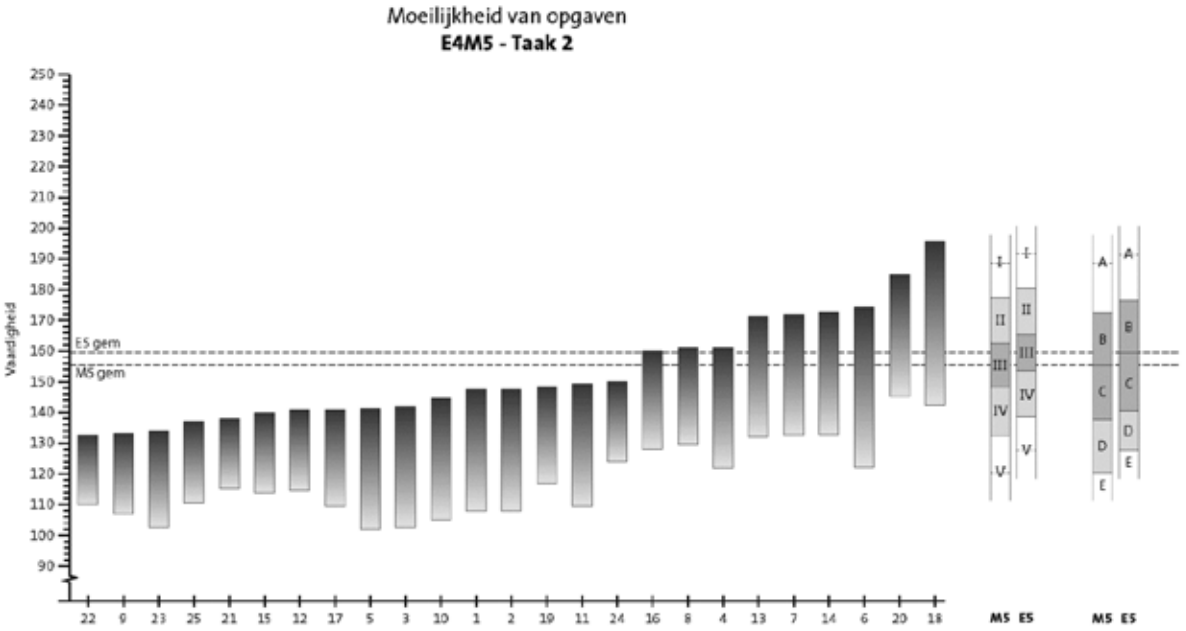
| Leesvaardigheid Zakelijke teksten | Niveau 1F | Niveau 2F |
|---|---|---|
| **Algemene omschrijving lezen zakelijke teksten** | Kan eenvoudige teksten lezen over alledaagse onderwerpen en over onderwerpen die aansluiten bij de leefwereld. | Kan teksten lezen over alledaagse onderwerpen, onderwerpen 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 informatiedichtheid; 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) route-beschrijvingen 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 reclameteksten, advertenties, folders, maar ook brochures van formele instanties of licht opiniërende artikelen uit tijdschriften. |
| **Kenmerken van de taakuitvoering:** | | |
| **Begrijpen** | Herkent specifieke informatie, wanneer naar één expliciet genoemde informatie-eenheid gevraagd wordt (letterlijk begrip).<br><br>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). | Kan de hoofdgedachte van een tekst weergeven en maakt onderscheid tussen hoofd- en bijzaken.<br>Herkent beeldspraak (letterlijk en figuurlijk taalgebruik).<br>Legt relaties tussen tekstdelen (inleiding, kern, slot) en teksten.<br>Ordent informatie (bijvoorbeeld op basis van signaalwoorden) voor een beter begrip. |
| **Interpreteren** | Kan informatie en meningen interpreten voor zover deze dicht bij de leerling staan. | Legt relaties tussen tekstuele informatie en meer algemene kennis.<br>Kan de bedoeling van tekstgedeeltes en/of specifieke formuleringen duiden.<br>Kan de bedoeling van de schrijver verwoorden. |
| **Evalueren** | Kan een oordeel over een tekst(deel) verwoorden. | Kan relaties tussen en binnen teksten evalueren en beoordelen. |

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

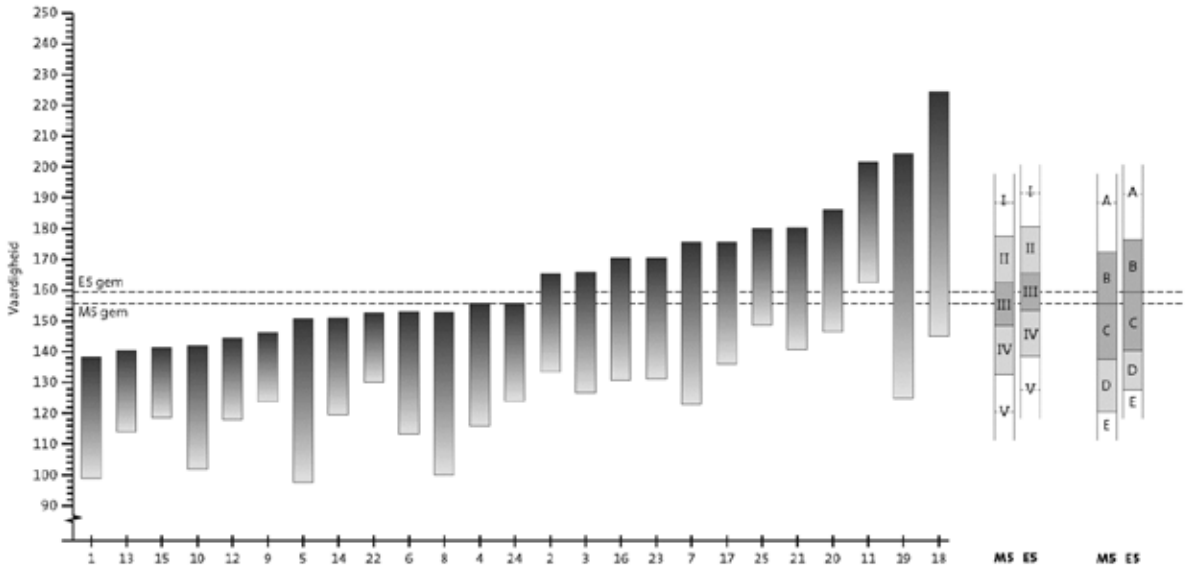
| Leesvaardigheid<br>Fictionele, narratieve en literaire teksten | Niveau 1F | Niveau 2F |
|---|---|---|
| **Algemene omschrijving**<br>**Lezen fictionele, narratieve en literaire teksten** | Kan jeugdliteratuur belevend lezen. | Kan eenvoudige adolescentenliteratuur herkennend lezen. |
| **Teksten:** | | |
| **Tekstkenmerken** | 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.<br>Poëzie en liedjes hebben meestal een verhalende inhoud en emotionele lading. |
| **Kenmerken van de taakuitvoering:** | | |
| **Begrijpen** | Herkent basale structuurelementen, zoals wisselingen van tijd en plaats,rijm en versvorm.<br>Kan meeleven met een personage en uitleggen hoe een personage zich voelt.<br>Kan gedichten en verhaalfragmenten parafraseren of samenvatten. | Herkent het genre.<br>Herkent letterlijk en figuurlijk taalgebruik. |
| **Interpreteren** | Kan relaties leggen tussen de tekst en de werkelijkheid.<br>Kan spannende, humoristische of dramatische passages in de tekst aanwijzen.<br>Herkent verschillende emoties in de tekst, zoals verdriet, boosheid en blijdschap. | Kan bepalen in welke mate personages en gebeurtenissen herkenbaar en realistisch zijn.<br>Kan personages typeren, zowel innerlijk als uiterlijk.<br>Kan het onderwerp in de tekst benoemen. |
| **Evalueren** | Evalueert de tekst met emotieve argumenten.<br><br>Kan met medeleerlingen leeservaringen uitwisselen.<br>Kan interesse in bepaalde fictievormen aangeven. | Evalueert de tekst ook met realistische argumenten en kan persoonlijke reacties toelichten met voorbeelden uit de tekst.<br>Kan met medeleerlingen leeservaringen uitwisselen en kan de interesse in bepaalde genres of onderwerpen motiveren. |

## Bijlage 2 Moeilijkheid van de opgaven

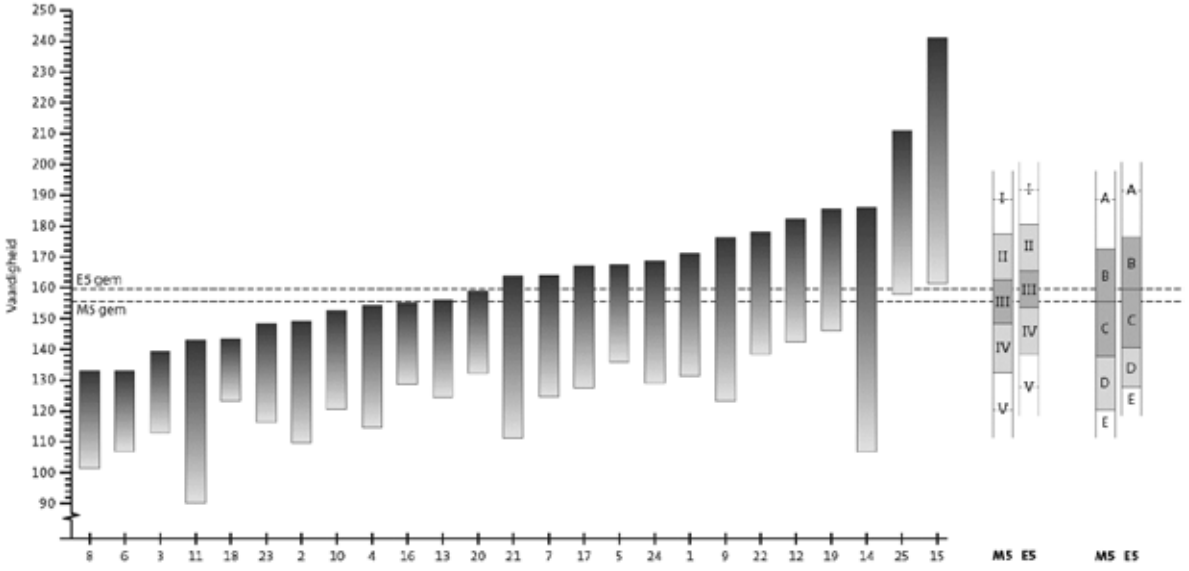


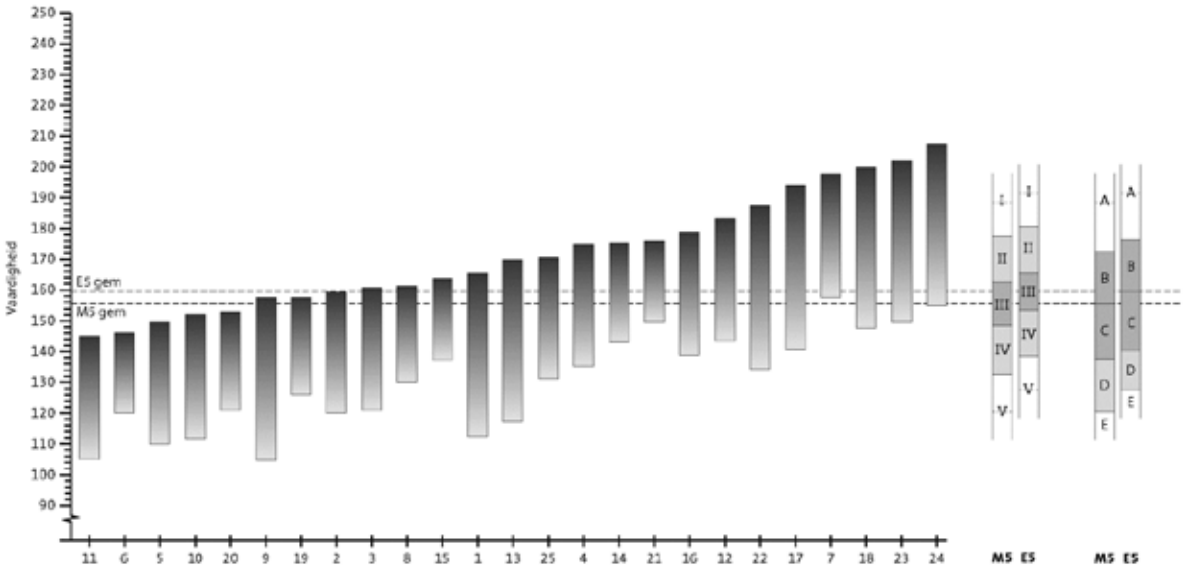
Moeilijkheid van opgaven
**M5 - Taak 1**



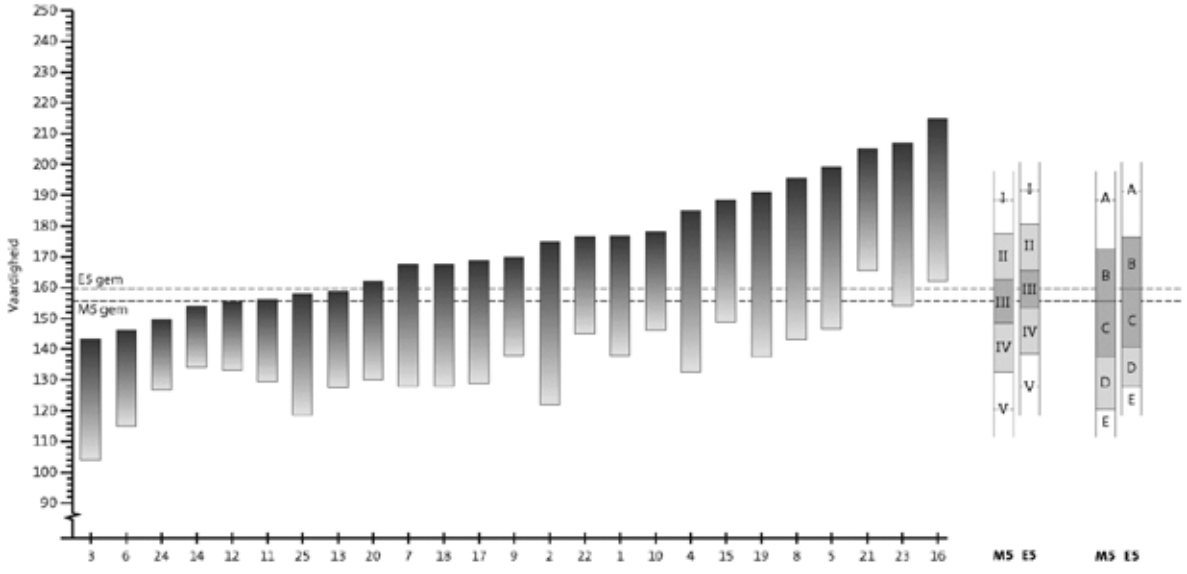

Moeilijkheid van opgaven
**M5 - Taak 2**



Moeilijkheid van opgaven
**E5 - Taak 1**




Moeilijkheid van opgaven
**E5 - Taak 2**



## Bijlage 3 Klassieke en IRT-indices van de opgaven in de E4M5-toets

| Volgnr boekje | p-waarde | RIT | Beta | Info |
|---|---|---|---|---|
| 1 | .58 | .34 | .13 | 1.988 |
| 2 | .92 | .33 | -.33 | 1.704 |
| 3 | .71 | .45 | .02 | 4.146 |
| 4 | .84 | .28 | -.35 | 1.164 |
| 5 | .70 | .40 | -.01 | 2.89 |
| 6 | .76 | .38 | -.10 | 2.537 |
| 7 | .84 | .40 | -.16 | 2.901 |
| 8 | .74 | .45 | -.02 | 3.909 |
| 9 | .82 | .29 | -.30 | 1.268 |
| 10 | .72 | .39 | -.03 | 2.808 |
| 11 | .55 | .34 | .17 | 2.012 |
| 12 | .64 | .47 | .11 | 4.628 |
| 13 | .80 | .42 | -.11 | 3.286 |
| 14 | .63 | .47 | .11 | 4.656 |
| 15 | .91 | .46 | -.18 | 4.027 |
| 16 | .75 | .38 | -.08 | 2.613 |
| 17 | .81 | .36 | -.18 | 2.174 |
| 18 | .78 | .48 | -.04 | 4.743 |
| 19 | .62 | .52 | .14 | 6.307 |
| 20 | .79 | .53 | -.02 | 5.974 |
| 21 | .86 | .51 | -.10 | 5.414 |
| 22 | .62 | .41 | .10 | 3.202 |
| 23 | .69 | .46 | .05 | 4.347 |
| 24 | .71 | .32 | -.09 | 1.693 |
| 25 | .85 | .44 | -.14 | 3.619 |
| 26 | .80 | .36 | -.17 | 2.229 |
| 27 | .80 | .36 | -.17 | 2.233 |
| 28 | .83 | .35 | -.21 | 2.02 |
| 29 | .73 | .39 | -.04 | 2.747 |
| 30 | .83 | .35 | -.22 | 1.993 |
| 31 | .69 | .33 | -.04 | 1.769 |
| 32 | .66 | .40 | .05 | 3.073 |
| 33 | .71 | .46 | .02 | 4.188 |
| 34 | .87 | .43 | -.18 | 3.265 |
| 35 | .82 | .36 | -.19 | 2.124 |
| 36 | .80 | .37 | -.15 | 2.295 |
| 37 | .83 | .46 | -.11 | 3.999 |
| 38 | .66 | .40 | .05 | 3.061 |
| 39 | .66 | .40 | .05 | 3.085 |
| 40 | .84 | .45 | -.12 | 3.919 |
| 41 | .72 | .45 | .01 | 4.104 |
| 42 | .83 | .41 | -.15 | 2.932 |
| 43 | .57 | .34 | .14 | 1.993 |
| 44 | .79 | .43 | -.09 | 3.441 |
| 45 | .57 | .41 | .16 | 3.329 |
| 46 | .85 | .49 | -.10 | 4.809 |
| 47 | .88 | .46 | -.15 | 4.084 |
| 48 | .87 | .38 | -.21 | 2.495 |
| 49 | .77 | .49 | -.03 | 4.888 |
| 50 | .85 | .44 | -.14 | 3.632 |

**Klassieke en IRT-indices van de opgaven in de M5- toets**

| Volgnr boekje | p-waarde | RIT | Beta | Info |
|---|---|---|---|---|
| 1 | .85 | .33 | -.25 | 1.872 |
| 2 | .68 | .46 | .06 | 4.384 |
| 3 | .70 | .40 | .00 | 2.895 |
| 4 | .76 | .38 | -.10 | 2.525 |
| 5 | .80 | .29 | -.26 | 1.354 |
| 6 | .78 | .37 | -.12 | 2.43 |
| 7 | .68 | .33 | -.04 | 1.78 |
| 8 | .79 | .30 | -.23 | 1.397 |
| 9 | .79 | .52 | -.03 | 5.858 |
| 10 | .83 | .34 | -.22 | 2.008 |
| 11 | .45 | .40 | .31 | 3.357 |
| 12 | .81 | .46 | -.08 | 4.34 |
| 13 | .84 | .45 | -.11 | 3.941 |
| 14 | .78 | .43 | -.07 | 3.577 |
| 15 | .83 | .50 | -.07 | 5.227 |
| 16 | .67 | .40 | .04 | 3.03 |
| 17 | .64 | .41 | .08 | 3.161 |
| 18 | .54 | .26 | .16 | 0.946 |
| 19 | .62 | .25 | -.02 | 0.897 |
| 20 | .56 | .41 | .17 | 3.342 |
| 21 | .60 | .41 | .12 | 3.257 |
| 22 | .74 | .54 | .03 | 6.68 |
| 23 | .67 | .40 | .04 | 3.032 |
| 24 | .75 | .44 | -.03 | 3.849 |
| 25 | .56 | .47 | .19 | 4.888 |
| 26 | .80 | .36 | -.15 | 2.304 |
| 27 | .84 | .45 | -.12 | 3.876 |
| 28 | .77 | .37 | -.10 | 2.497 |
| 29 | .67 | .40 | .04 | 3.05 |
| 30 | .66 | .46 | .08 | 4.517 |
| 31 | .87 | .42 | -.17 | 3.311 |
| 32 | .71 | .39 | -.02 | 2.848 |
| 33 | .87 | .37 | -.22 | 2.462 |
| 34 | .68 | .33 | -.03 | 1.795 |
| 35 | .77 | .44 | -.05 | 3.679 |
| 36 | .83 | .28 | -.32 | 1.217 |
| 37 | .59 | .41 | .14 | 3.292 |
| 38 | .74 | .44 | -.02 | 3.901 |
| 39 | .69 | .24 | -.17 | 0.821 |
| 40 | .47 | .25 | .31 | 0.949 |
| 41 | .73 | .50 | .02 | 5.36 |
| 42 | .69 | .40 | .01 | 2.938 |
| 43 | .81 | .54 | -.03 | 6.871 |
| 44 | .56 | .41 | .17 | 3.339 |
| 45 | .70 | .51 | .05 | 5.671 |
| 46 | .74 | .31 | -.13 | 1.602 |
| 47 | .62 | .41 | .10 | 3.22 |
| 48 | .79 | .42 | -.09 | 3.43 |
| 49 | .68 | .40 | .02 | 2.982 |
| 50 | .48 | .34 | .28 | 2.029 |

**Klassieke en IRT-indices van de opgaven in de E5-toets**

| Volgnr boekje | p-waarde | RIT | Beta | Info |
|---|---|---|---|---|
| 1 | .76 | .29 | -.12 | 1.548 |
| 2 | .77 | .36 | -.06 | 2.53 |
| 3 | .76 | .36 | -.05 | 2.568 |
| 4 | .68 | .38 | .07 | 3.063 |
| 5 | .82 | .33 | -.15 | 2.132 |
| 6 | .83 | .43 | -.06 | 4.073 |
| 7 | .51 | .39 | .27 | 3.444 |
| 8 | .74 | .42 | .03 | 3.992 |
| 9 | .79 | .28 | -.19 | 1.403 |
| 10 | .81 | .34 | -.13 | 2.23 |
| 11 | .84 | .32 | -.19 | 1.964 |
| 12 | .62 | .39 | .15 | 3.276 |
| 13 | .73 | .30 | -.09 | 1.629 |
| 14 | .64 | .45 | .14 | 4.707 |
| 15 | .71 | .48 | .09 | 5.792 |
| 16 | .65 | .39 | .11 | 3.167 |
| 17 | .61 | .32 | .12 | 1.963 |
| 18 | .57 | .33 | .18 | 2.016 |
| 19 | .77 | .41 | -.01 | 3.75 |
| 20 | .80 | .40 | -.05 | 3.428 |
| 21 | .60 | .50 | .20 | 6.574 |
| 22 | .64 | .32 | .07 | 1.894 |
| 23 | .56 | .33 | .20 | 2.026 |
| 24 | .53 | .33 | .24 | 2.048 |
| 25 | .70 | .38 | .04 | 2.94 |
| 26 | .66 | .39 | .09 | 3.129 |
| 27 | .71 | .31 | -.04 | 1.713 |
| 28 | .85 | .32 | -.20 | 1.9 |
| 29 | .66 | .32 | .05 | 1.865 |
| 30 | .58 | .33 | .17 | 2.01 |
| 31 | .83 | .38 | -.11 | 2.992 |
| 32 | .72 | .37 | .01 | 2.824 |
| 33 | .60 | .32 | .14 | 1.982 |
| 34 | .68 | .44 | .10 | 4.474 |
| 35 | .62 | .45 | .17 | 4.815 |
| 36 | .77 | .46 | .03 | 5.087 |
| 37 | .76 | .51 | .06 | 6.627 |
| 38 | .76 | .42 | .01 | 3.844 |
| 39 | .77 | .54 | .06 | 8.001 |
| 40 | .58 | .39 | .19 | 3.366 |
| 41 | .48 | .32 | .31 | 2.052 |
| 42 | .71 | .37 | .02 | 2.869 |
| 43 | .72 | .37 | .01 | 2.838 |
| 44 | .63 | .32 | .10 | 1.93 |
| 45 | .74 | .43 | .03 | 4.034 |
| 46 | .45 | .39 | .34 | 3.42 |
| 47 | .62 | .45 | .16 | 4.774 |
| 48 | .53 | .33 | .24 | 2.046 |
| 49 | .81 | .48 | .00 | 5.742 |
| 50 | .78 | .35 | -.07 | 2.478 |

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