Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3

Marieke Tomesen, Jasper Wouda, Astrid Mols en Linda Horsels

zeker weten

Wetenschappelijke verantwoording van de LVS-toetsen

Spelling 3.0 voor groep 3

Marieke Tomesen
Jasper Wouda
Astrid Mols
Linda Horsels

Cito
Arnhem, 2015

Inhoud

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de papieren LVS-toetsen Spelling 3.0 voor groep 3. De toetsen Spelling 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 primair onderwijs. Het betreft papieren en digitale toetsen[1] voor alle leerjaren.
Dit jaar wordt ook de wetenschappelijke verantwoording voor de papieren toetsen Spelling 3.0 groep 4 uitgebracht. Te zijner tijd zullen ook de wetenschappelijke verantwoordingen voor de digitale toetsen worden uitgebracht, evenals de wetenschappelijke verantwoordingen met de gegevens van de (nog te verschijnen) toetsen Spelling 3.0 voor de groepen 5 t/m 8.

Deze verantwoording biedt tezamen met de inhoud van het toetspakket Spelling 3.0 voor groep 3 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 Spelling 3.0 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 Spelling 3.0 voor groep 3. De kwaliteit van het toetsmateriaal en de handleiding is te bepalen door kennis te nemen van de inhoud van het toetspakket.

[1] De digitale toetsen voor Spelling 3.0 voor groep 3 worden in januari 2016 uitgebracht.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

Bij het spellen wordt de gesproken taal omgezet in geschreven taal. Om woorden correct te schrijven, moeten leerlingen spellingregels en/of spellingstrategieën kunnen toepassen.
De toetsen in de toetspakketten Spelling 3.0 van het Cito Volgsysteem primair en speciaal onderwijs zijn bedoeld om vast te stellen hoe goed een leerling kan spellen en hoe de spellingvaardigheid van de leerling zich in de loop van de jaren ontwikkelt.
Het vaststellen van de spellingvaardigheid gebeurt door de leerling woorden te laten opschrijven (actieve spelling). De spellingregels zelf worden niet expliciet bevraagd. De leerling laat indirect zien dat hij of zij de spellingregels beheerst door de gevraagde woorden correct te schrijven (zie verder paragraaf 2.4.1). In de toetsen voor groep 3 wordt enkel de spellingvaardigheid van niet-werkwoorden getoetst. Werkwoordspelling komt in het onderwijs namelijk pas later aan bod.

2.2 Doelgroep

De toetsen Spelling 3.0 voor groep 3 van het Cito Volgsysteem primair en speciaal onderwijs zijn bestemd voor leerlingen in groep 3 van het basisonderwijs. De toetsen zijn ook geschikt voor leerlingen in het speciaal basisonderwijs en in 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 (basis)onderwijs een correcte referentiegroep.

Voor de toetsen van groep 3 zijn zowel voor ‘midden leerjaar’ (half januari/half februari) als voor ‘einde leerjaar’ (juni) populatieparameters bepaald. De toetsen kunnen desgewenst ook op andere momenten in het schooljaar worden afgenomen, maar dat maakt het moeilijker om uitspraken te doen over het niveau van de leerling ten opzichte van andere leerlingen in Nederland.
De toetsen Spelling 3.0 voor groep 3 kunnen ook gebruikt worden voor leerlingen in andere leerjaren die werken op het niveau van groep 3. In de handleiding is toegelicht hoe dit toetsen op maat, met behulp van vaardigheidsscores, in zijn werk gaat. Voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften zijn in de handleiding extra aanwijzingen opgenomen. Voor deze leerlingen zijn alternatieve rapportageformulieren ontwikkeld en zijn extra toetsen beschikbaar. Door het toevoegen van deze extra toetsen (tussentoetsen) beslaan opeenvolgende toetsen kleinere leerstappen. Zo is er naast een toets voor het functioneringsniveau M3 (medio groep 3) en E3 (eind groep 3) ook een toets voor het functioneringsniveau M3E3. Deze laatste is wat moeilijker dan M3 en wat gemakkelijker dan E3. Aan een leerling die zich minder snel ontwikkelt in spellingvaardigheid, kan aan het eind van groep 3 dus de toets M3E3 voorgelegd worden. Deze leerling hoeft zo niet een te moeilijke toets (E3) te maken, maar ook niet twee keer dezelfde toets (M3). Voor de extra toetsen zijn de parameters van de reguliere afnamemomenten gebruikt. Het uitgangspunt is dat de normering van de extra toets gebaseerd is op de populatieparameters van een hoger, regulier afnamemoment omdat een extra toets een gemakkelijkere variant is van de opvolgende reguliere toets. Voor de normering van de extra toets M3E3 zijn de populatieparameters gebruikt van afnamemoment einde groep 3.

Het is pas zinvol om de toets M3 af te nemen als alle letters tijdens de spellinglessen aangeboden zijn. Voor alle toetsen geldt dat ze niet geschikt zijn voor leerlingen met een (tijdelijk) beperkt gehoor; in dit geval is niet bekend of een slechte prestatie toe te schrijven is aan de spellingvaardigheid of aan de gehoorproblemen. In de handleiding geven we het advies om bij vermoeden van een tijdelijk gehoorverlies de toets op een later moment af te nemen.

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 Spelling 3.0 uit het Cito Volgsysteem hebben twee doelen: niveaubepaling en progressiebepaling. Tevens wordt in de toetsen Spelling de mogelijkheid geboden de door de leerling gemaakte fouten te analyseren met het oog op het aanbieden van gerichte remediëring. Deze 'signalering' staat geheel los van de niveau- en progressiebepaling en is als zodanig ook niet meegenomen in de kalibratie- en normeringsonderzoeken.

Niveaubepaling
De toetsen Spelling geven de leerkracht informatie over het niveau van de spellingvaardigheid van zijn leerlingen, individueel en als groep. Iedere behaalde vaardigheidsscore kan normgericht geïnterpreteerd worden op basis van de vaardigheidsverdeling in een adequate 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 in vijf niveaugroepen verdeeld. Er is sprake van twee indelingen. De eerste indeling, gebaseerd op de niveaugroepen I tot en met V, gaat uit van vijf groepen van ieder 20%. Bij deze indeling worden op de registratie-overzichten de laagste 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 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 deze indeling in de handleiding steeds als eerste genoemd.

[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 spelling, 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 Spelling 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 Spelling 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 vaardigheidsniveaus (die zijn afgebakend door percentielpunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk zoals "Meriam heeft op afnamemoment einde leerjaar 3 vaardigheidsniveau IV behaald". Voor de leerkracht (en voor Meriam en haar ouders) bevat deze uitspraak waardevolle informatie. De leerkracht kan op basis hiervan bijvoorbeeld besluiten om Meriam extra lesstof aan te bieden.

Ad b.
Voor het vergelijken ('volgen') van een leerling op twee verschillende tijdstippen komen twee methodes in aanmerking. Bij de eerste methode worden de **vaardigheidsniveaus** op de twee tijdstippen vergeleken: "op tijdstip E3 had Meriam vaardigheidsniveau IV en op tijdstip M4 was het vaardigheidsniveau V". Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 188, bijvoorbeeld, op tijdstip E3 en vaardigheidsscore 201 op tijdstip M4. Ook hier geldt, net als bij het vergelijken van prestaties met die van andere leerlingen, dat bij eventuele verdere acties van de leerkracht ook andere aspecten moeten worden betrokken.

Bij alle vergelijkingen die mogelijk zijn, zowel die ad a. als die ad b., dienen uitspraken over leerlingen te worden gerelativeerd. In de handleiding is meer informatie te vinden over de wijze waarop de gebruiker dit kan doen. Hieronder gaan we vooral in op het belang van de (on)betrouwbaarheid van de afgenomen toetsen hierbij.
Voor elke toets geldt dat de vaardigheidsscore die bij een toetsresultaat van een leerling hoort, behept is met een meetfout. Als we rekening houden met die meetfout dan zou het best zo kunnen zijn dat Meriam 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: Meriam 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 Spelling E3 behaalde Wout een vaardigheidsscore van 202 met een 67% betrouwbaarheidsinterval van 184-220. Bij de afname M4 behaalde Wout een vaardigheidsscore van 236; het bijbehorende betrouwbaarheidsinterval daarbij is 222-250. Aangezien de betrouwbaarheids-intervallen niet overlappen kunnen we zeggen dat Wouts vaardigheid is toegenomen.

Conclusie
De vaardigheidsgroei voor Spelling voltrekt zich langzaam in de tijd. De verschillen tussen vaardigheids-scores op achtereenvolgende meettijdstippen zijn betrekkelijk klein, al is de gemiddelde vaardigheids-toename in groep 3 en groep 4 wat groter dan in de hogere leerjaren. Bovendien is er sprake van meetfouten. De verschillen in vaardigheidsgroei moeten tegen de achtergrond van die meetfouten worden

geïnterpreteerd. Dit betekent dat men weliswaar uitspraken kan doen over de vaardigheidsgroei van een leerling, maar dat deze uitspraken met voorzichtigheid dienen te worden gehanteerd. Dit geldt ook wanneer men de progressie van een leerling volgt in termen van vaardigheidsniveaus of een vergelijking maakt met andere leerlingen in termen van 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.

'Signalering' via foutenanalyse
Als veel leerlingen fouten maken bij een bepaalde spellingcategorie, kan dat een signaal zijn dat het aangeboden onderwijs in die categorie ontoereikend is geweest. Dat hoeft niet direct alarmerend te zijn; misschien komt de betreffende spellingcategorie in de gebruikte methode pas op een later tijdstip aan de orde. Als de categorie daarentegen al wel is behandeld, kunnen de tegenvallende prestaties voor de leerkracht een reden zijn om nogmaals expliciet en voor de hele groep op de bij die categorie behorende spellingregels terug te komen. Door het invullen van een analyseformulier of het invoeren van de leerling-antwoorden in het Computerprogramma LOVS kan de leerkracht nagaan met welke spellingcategorieën een of meerdere leerlingen problemen hadden in de toets Spelling.
Individuele leerlingen die blijk geven van onvoldoende beheersing van een of meerdere categorieën zullen wellicht baat hebben bij extra instructie en gerichte oefeningen. Omdat het aantal opgaven per categorie in een toets Spelling beperkt is (er zijn veel categorieën en de toets mag niet te lang worden), kan niet worden uitgesloten dat de leerling bij toeval juist de opgaven uit deze categorie fout heeft beantwoord. Om meer zekerheid te verkrijgen over de beheersing van de categorie door deze leerling, kan de leerkracht gebruikmaken van een controledictee. Controledictees zijn opgenomen in de handleiding van de toetsen. Elk controledictee bevat tien opgaven uit één bepaalde categorie. Als de leerling zeven of minder opgaven goed heeft van het controledictee, dan heeft de categorie voor die leerling extra aandacht nodig.
De leerkracht kan deze leerling vervolgens aanvullende instructie en/of oefenmateriaal aanbieden, bij voorkeur uit de eigen methode (zie ook hoofdstuk 4 van de handleiding).

Er is één uitzondering: na afname van de toets M3 zijn controledictees niet nodig. De toets M3 bevat slechts twee categorieën, namelijk de mkm-woorden (categorie 1) en de mmkm- en mkmm-woorden (categorie 2) die respectievelijk 14 en 26 woorden bevatten. Op basis van deze aantallen heeft een leerkracht voldoende informatie over het al dan niet beheersen van een categorie.

Er is geen kwalitatief of kwantitatief onderzoek gedaan naar het adequaat functioneren van de fouten-analyse en de 'doorverwijzing' via de controledictees. De signalering via foutenanalyse heeft dan ook geen wetenschappelijke status of pretentie. Haar enige functie is leerkrachten die gericht extra ondersteuning willen bieden aan leerlingen die moeite hebben met het correct spellen van bepaalde woorden een handreiking doen.

2.4 Theoretische inkadering

2.4.1 Inhoudelijk

Wat is spelling?
Spelling is een ondersteunende taalvaardigheid die instrumenteel is voor schrijven. Het is een aspect van codeervaardigheid, waarbij het gaat om de correcte schrijfwijze van woorden. Ondersteunende taalvaardigheden hebben tot doel de zogeheten functionele taalactiviteiten – activiteiten waarbij de taal als communicatiemiddel fungeert, zoals het schrijven van een briefje – beter te kunnen uitvoeren.
Voor een beschrijving van het begrip spelling hanteren we de definitie van De Schryver & Neijt (2005). Zij omschrijven spelling als '…een systeem van regels met behulp waarvan we een bepaalde gesproken taal schriftelijk weergeven' (De Schryver & Neijt, 2005, p. 15). De laatste 'versie' van de spelling van het Nederlands is in 2005 vastgelegd in de Woordenlijst Nederlandse Taal, ook wel 'het Groene Boekje' genoemd. Het gebruik van deze spelling is verplicht binnen het onderwijs.

Binnen de Nederlandse spelling is er niet altijd sprake van een één-op-één relatie tussen klank en letterteken. Het Nederlands kent circa 40 spraakklanken, ook wel fonemen genoemd. Een foneem is de kleinste onderscheidende klankeenheid in een taal. De woorden *bot* en *boot* bestaan beide uit drie fonemen, maar verschillen van elkaar als gevolg van een verschillend klinkerfoneem. Het alfabet heeft maar 26 letters. Dit betekent dat dezelfde letters voor verschillende klanken gebruikt moeten worden: d**e**ling, b**e**l, raf**e**l. Maar andersom wordt een klank ook door verschillende tekens weergegeven: p**ij**l, p**ei**l. Verder bevat de Nederlands spelling hulptekens om de klank van (groepen) letters duidelijker weer geven, bijvoorbeeld het accent in café en het koppelteken in co-ouder.

De spelling van de Nederlandse taal is gebaseerd op het basisbeginsel van de uitspraak van de fonemen in het Standaardnederlands. Dit wordt ook wel het ***fonologisch*** principe genoemd: een woord wordt gespeld met de fonemen die hoorbaar zijn in de standaarduitspraak van het losse woord. Hierbij worden kleine uitspraaknuances die ontstaan door persoonsgebonden of streekgebonden verschillen of door klanken in de omgeving van het woord genegeerd. Bijvoorbeeld: de /z/ in 'ik **z**et' klinkt als een /s/ in tegenstelling tot de /z/ in '**z**et ik'. Het fonologisch principe is het basisprincipe, maar er zijn allerlei uitzonderingen op deze hoofdregel. Die uitzonderingen zijn veelal niet willekeurig, maar hebben weer te maken met andere principes of regels.

Het ***morfologisch*** principe doorkruist het fonologisch principe en gaat uit van de morfologische structuur van een woord. Een morfeem is een betekenisdragend woorddeel. Het kan zowel om gehele woorden gaan als om voor- of achtervoegsels, zoals 'on-' en '-heid'. Bij het morfologisch principe onderscheiden we twee regels: de regel van de gelijkvormigheid en de regel van de overeenkomst. De regel van de gelijkvormigheid houdt in dat we een woord of een voor- of achtervoegsel steeds op dezelfde manier schrijven. Bijvoorbeeld: we schrijven 'hond' omdat we in het meervoud 'honden' een /d/ horen. De regel van de overeenkomst houdt in dat de opbouw van een woord duidelijk wordt in de spelling. Bijvoorbeeld: een woord als 'breedte' wordt zo gespeld, en niet als 'brete', omdat in 'breedte' de morfologische structuur van het woord (breed-breedte) zichtbaar is. Het morfologisch principe geldt zolang het niet met de uitspraak in conflict is. Bijvoorbeeld: je spelt 'bloem**p**je' omdat je het zo hoort, en niet 'bloemtje'. Echter, bij deze twee regels gaat het om hetzelfde principe: morfemen worden zo veel mogelijk op gelijke wijze gespeld.

Het ***etymologisch*** principe houdt in dat als er meerdere mogelijkheden zijn om een woord te schrijven, de schrijfwijze wordt gekozen zoals deze zich in het verleden heeft gevormd. Er is hier geen sprake van een regel, maar van kennis die we ons per woord eigen moeten maken. Voorbeelden hiervan zijn de lettercombinaties ou/au en ei/ij. Vroeger, en in sommige dialecten nog steeds, gaven deze verschillende lettercombinaties verschillende klanken weer, maar nu zullen we in de meeste gevallen de spelling van dergelijke woorden gewoon uit het hoofd moeten leren.

De hiervoor genoemde principes gaan deels terug tot de 16e eeuw, en zijn door De Vries en Te Winkel in 1866 in een regeling ‘De grondbeginselen der Nederlandsche spelling’ vastgelegd. Daarnaast wordt de spelling van het Nederlands in belangrijke mate bepaald door twee zeer algemene regels: de regel voor het verdubbelen van medeklinkerletters en de regel voor het verenkelen van klinkerletters. De regel voor verenkeling schrijft voor dat als een syllabe eindigt op een lange klank we maar één letter schrijven, bijvoorbeeld in 'boten'. De verdubbelingsregel houdt in dat als een syllabe eindigt op een korte klank, de medeklinker die daarop volgt verdubbeld wordt, bijvoorbeeld in 'botten'. Ook op deze regels zijn echter weer uitzonderingen.
De terugleesbaarheid van heel wat woorden wordt mede bepaald door het onderscheid tussen enkele en dubbele medeklinker of klinker; de lezer zou daardoor zo veel mogelijk op grond van de geschreven vorm van het woord opnieuw bij het basisbeginsel van de uitspraak van het losse woord moeten uitkomen en niet bij een verkeerde uitspraak.

Ten slotte zijn er nog tal van bijzondere regels over allerlei onderdelen van het spellingsysteem.
Een voorbeeld hiervan is dat de ‘lange’ klinker /o/ aan het einde van een woord standaard met een *o* wordt

geschreven, zoals in *auto* en *zo,* met uitzondering van enkele woorden als *shampoo* (zie onder meer Het Groene Boekje, 2005; Nederlandse Taalunie; 2009; Van Bon, 1993).

Spellingstrategieën

De manieren die spellers gebruiken om tot de juiste schrijfwijze te komen (spellingstrategieën) worden vaak uitgesplitst in een directe strategie en indirecte strategieën (Bonset & Hoogeveen, 2009; Van der Beek & Paus, 2011). Als een speller de directe strategie gebruikt, schrijft hij een woord op zonder erbij na te denken. Het spellen is dan geautomatiseerd. Indirecte strategieën vinden plaats als je bij het spellen een bepaalde denkhandeling toepast. Huizenga onderscheidt vijf indirecte spellingstrategieën: de fonologische strategie, de woordbeeldstrategie, de regelstrategie, de analogiestrategie en de hulpstrategie.
De ***fonologische*** strategie houdt in dat je bij het spellen uitgaat van de klanken of klankgroepen waaruit een woord bestaat. Er zijn twee verschillende fonologische strategieën: de elementaire spellinghandeling, waarbij een woord wordt ontleed in fonemen, en de klankclusterstrategie, waarbij een woord wordt ontleed in klankgroepen. De elementaire spellinghandeling is normaal gesproken de eerste die een kind leert (voor het Nederlands). Ze is bruikbaar zolang een leerling alleen klankzuivere woorden moet schrijven (in het basisonderwijs wordt vaak gesproken van 'luisterwoorden'). De klankclusterstrategie is bruikbaar voor het schrijven van klankgroepen die altijd door dezelfde lettercombinatie worden weergegeven, bijvoorbeeld -ooi of -uw. In het basisonderwijs wordt dit wel aangeduid met de term 'luisterwoorden met speciale klankgroepen'. Deze strategie is voor leerlingen wat lastiger dan de elementaire spellinghandeling.
De ***woordbeeldstrategie*** houdt in dat je een woord correct schrijft door een beroep te doen op het woordgeheugen. Deze strategie is vooral bruikbaar bij leenwoorden of woorden waarvan de schrijfwijze moet worden ingeprent, bijvoorbeeld woorden met -ou- of -au-. In het basisonderwijs duidt men dergelijke woorden wel aan met de termen 'weetwoorden' of 'afspraakwoorden'.
De ***regelstrategie*** wordt gebruikt als je bij het schrijven van een woord een spellingregel toepast. Voorbeelden daarvan zijn de verenkelingsregel en de verdubbelingsregel, maar ook regels als 'Hoor je op het einde /-ies/, dan schrijf je -isch.' Op de meeste spellingregels zijn weer uitzonderingen en dat maakt deze strategie lastig. In het basisonderwijs gebruikt men wel de term 'regelwoorden'.
Bij de ***analogiestrategie*** schrijf je een woord door het te vergelijken met een ander woord. Die vergelijking kan gebaseerd zijn op overeenkomst in klank (bijvoorbeeld 'komen' en 'dromen'), maar ook op overeenkomst in betekenis (bijvoorbeeld 'vertrouwelijk' en 'trouwen'). De strategie leidt niet altijd tot het juiste resultaat, omdat de gemaakte vergelijking niet altijd opgaat (bijvoorbeeld 'hond', 'wond', 'lont'*)*. In het basisonderwijs worden de termen 'voorbeeldwoorden' of 'net-als woorden' gehanteerd.
De ***hulpstrategie*** houdt in dat je ezelsbruggetjes of hulpregels gebruikt om te onthouden hoe een woord gespeld moet worden. Deze kunnen zelfbedacht zijn, maar ook aangeleerd in het onderwijs.

Woorden kunnen vaak met verschillende strategieën goed geschreven worden. Zwakke spellers schrijven vaak letterlijk op wat ze horen (fonologische strategie) (Gijsel, Scheltinga, Van Druenen & Verhoeven, 2011a). Een volwassen speller zal voor veelvoorkomende, gemakkelijke woorden zoals 'school' waarschijnlijk de directe strategie gebruiken, maar hij kan ook de woordbeeldstrategie gebruiken.

Spelling in het basisonderwijs

Vanaf het moment dat een kind op school leert lezen en schrijven, wordt er aandacht besteed aan spelling. Daarbij wordt onderscheid gemaakt tussen aanvankelijk spellen (groep 3) en voortgezet spellen (vanaf groep 4). In de fase van het aanvankelijk spellen leert een leerling klankzuivere eenlettergrepige woorden schrijven. In de fase van het voortgezet spellen komen niet-klankzuivere meerlettergrepige woorden aan bod (Gijsel, Scheltinga, Van Druenen & Verhoeven, 2011a, 2011b; Bonset & Hoogeveen, 2009). Voor het correct spellen van woorden zijn vele strategieën mogelijk. Een kind dat leert spellen, moet deze spellingstrategieën aanleren en op elkaar afstemmen. Binnen het onderwijs wordt steeds meer rekening gehouden met het feit dat leerlingen gebruik kunnen maken van verschillende strategieën om een woord correct te spellen. In de recente methoden komen dan ook de hierboven genoemde strategieën, zij het soms onder een andere naam, steeds weer terug.

Om te bepalen welke leerstof aan bod moet komen in het spellingonderwijs worden in de handleidingen van taalmethoden meestal de volgende criteria genoemd:

1 de frequentie van woorden;
2 de moeilijkheid van woorden;
3 de indeling in spellingcategorieën.

Ad 1. In de huidige taalmethoden wordt de spelling behandeld van de 4000 meest frequente woorden (bij benadering) die voorkomen in Nederlandse teksten. Dit is een efficiënte aanpak, want als leerlingen deze woorden correct kunnen spellen, zullen zij al veel teksten vrijwel foutloos schrijven. Leerlingen leren om in geval van minder bekende woorden het woordenboek te raadplegen.

Ad 2. Ook de moeilijkheid van woorden is een criterium. De meest frequente woorden zijn vaak eenvoudig om te spellen. In het spellingonderwijs komen daarom (in de hogere leerjaren) ook woorden aan bod die minder frequent voorkomen en die vaak fout gespeld worden. Aan deze woorden wordt aandacht besteed omdat het bij het schrijven niet handig is deze woorden steeds op te moeten zoeken. Het gaat dan om woorden als: museum, enigszins, directie, chauffeur.

Ad 3. Tot slot wordt voor de ordening van de leerstof verder uitgegaan van een indeling in spelling-categorieën, groepen woorden met dezelfde spellingmoeilijkheid, dan wel -problemen. Deze categorieën zijn een hulpmiddel voor leerkrachten en methodemakers om de leerstof te ordenen en de spellingvaardigheid te diagnosticeren, maar ook om de lesstof voor leerlingen te structureren. De volgorde waarin de verschillende categorieën aan bod komen in de verschillende spellingmethoden en leerjaren is over het algemeen vergelijkbaar (Huizenga, 2010).

In het ‘Referentiekader taal en rekenen’ (Expertgroep Doorlopende Leerlijnen taal en rekenen, 2009a; Van der Beek & Paus, 2011) worden de spellingcategorieën ingedeeld in vijf klassen om de moeilijkheid van spelling te ordenen. Deze taalkundige indeling wordt volgens Kleijnen (1997; 2004) en Schijf (2009) ook gebruikt bij het diagnosticeren van spellingvaardigheid. De vijf klassen, gebaseerd op een analyse van het Nederlandse taalsysteem, worden hierna beschreven.

Alfabetische spelling is gebaseerd op een een-op-een-relatie tussen klank en teken. Het schrijven van klankzuivere woorden behoort tot de elementaire spelhandeling. In groep 3 leren leerlingen vooral klankzuivere woorden te schrijven. Woorden als ‘tak’, ‘krant’ en ‘ziek’ worden dan voor het eerst aangeboden.
Aan de basis van **orthografische** spelling liggen afspraken over de schrijfwijzen van (groepen) woorden, waarbij er geen automatische klank-letteromzetting plaatsvindt. Deze woorden volgen geen regels maar moeten door leerlingen ingeprent of geleerd worden, zoals woorden met -ieuw of een verdubbeling bij de meervoudsvorm.
De **lexicaal-morfologische** spelling is spelling op basis van de opbouw van het woord, los van de grammaticale context. De speller moet inzicht hebben in de subdelen van het woord, ook wel verwante woorddelen genoemd. Doordat deze delen hetzelfde worden geschreven komt de leerling tot een correcte spelling. Het volstaat om naar het woord zelf te kijken. Woorden als ondiep (voor-achtervoegsel), boompje (verkleinwoord) en tuindeur (samenstelling) komen vanaf groep 4 aan bod.
Binnen het onderwijs komen kinderen ook in aanraking met de grammaticale context om een woord goed te kunnen spellen. Dit noemen we **morfologische spelling op syntactische basis**. Het betreft vooral werkwoordvormen als ‘ik word/hij wordt’, maar ook woorden als ‘alle(n)’ en ‘enkele(n)’. Dit vraagt om een hogere spellingvaardigheid. Leerlingen komen hier pas mee in aanraking vanaf groep 6.
Zowel in de hoge als in de lage groepen krijgen leerlingen **logografische** spelling aangeboden. Logografische spelling is gebaseerd op vaststaande combinaties zonder regelvorming, ofwel woorden met een specifieke schrijfwijze. Hier gaat het om relatief eenvoudige woorden als ‘trein’ en ‘lijst’, maar ook om leenwoorden, zoals bijvoorbeeld ‘trottoir’ en ‘team’.

De indeling is zeer globaal en is niet afgebakend per leerjaar. Al beginnen alle methoden natuurlijk wel met de klankzuivere woorden en wordt er in de meeste methoden in groep 6 een begin gemaakt met de

werkwoordspelling. Categorieën die op een bepaald moment nieuw worden aangeboden, worden in de daaropvolgende leerjaren steeds herhaald.

Wettelijke basis voor het spellingonderwijs
De wettelijke basis voor het onderwijs in spelling is vastgelegd in het 'Referentiekader taal en rekenen' (Expertgroep Doorlopende Leerlijnen, 2009a). Hierin staat beschreven wat kinderen op verschillende momenten in hun schoolloopbaan op het gebied van taal en rekenen moeten kennen en kunnen. Het referentiekader onderscheidt voor taal vier domeinen: Mondelinge taalvaardigheid, Lezen, Schrijven en Begrippenlijst en taalverzorging. Er zijn voor deze domeinen vier niveaus onderscheiden. Die niveaus zijn de fundamentele niveaus (F) genoemd. Het fundamenteel niveau 1 (niveau 1F) voor het eind van het primair en speciaal onderwijs en het praktijkonderwijs, niveau 2F voor mbo 1, 2, 3 en vmbo, niveau 3F voor mbo 4 en eind havo en ten slotte niveau 4F voor eind vwo. Leerlingen die een fundamenteel niveau hebben behaald, krijgen meer aangeboden: ze gaan op weg naar het volgende niveau, het zogenoemde streef-niveau. Het streefniveau 1S voor het primair en speciaal onderwijs staat gelijk aan niveau 2F. Voor de leerlingen die het fundamenteel niveau 1F op het eind van de basisschool niet halen, biedt de leerkracht adequate leerstof aan, aansluitend op de mogelijkheden van de leerlingen. Het geheel aan beschrijvingen wordt aangeduid met 'het referentiekader' en is vastgelegd in de Wet referentieniveaus Nederlandse taal en rekenen die op 1 augustus 2010 van kracht is geworden.
Voor spelling is het domein 'Begrippenlijst en taalverzorging' van belang, en dan uitsluitend het onderdeel taalverzorging. De *begrippenlijst* omvat begrippen en concepten die leerlingen moeten kennen en kunnen hanteren om over taal en taalverschijnselen te kunnen denken en spreken. Bij *taalverzorging* gaat het om kennis die in dienst staat van een verzorgde schriftelijke taalproductie, en in het referentiekader wordt dat beperkt tot regels voor spelling, interpunctie en het gebruik van hoofdletters. De inhoud van het domein taalverzorging sluit aan bij twee kerndoelen Nederlandse taal voor het basisonderwijs. In kerndoel 8 staat dat leerlingen aandacht leren besteden aan correcte spelling, in kerndoel 11 dat ze regels leren voor het spellen van werkwoorden, voor andere woorden dan werkwoorden, en voor het gebruik van leestekens.
De niveaus 1F en 2F geven een eindpunt aan. In de publicatie 'Leerstoflijnen begrippenlijst en taal-verzorging beschreven' (Van der Beek & Paus, 2011) is aangegeven langs welke weg de eindniveaus 1F en 1S/2F te bereiken zijn. Deze publicatie geeft een antwoord op de vraag hoe de opbouw van de leerstoflijnen voor Begrippenlijst en taalverzorging eruit kan zien. Deze leerstoflijnen kunnen worden gebruikt voor de planning en opbouw van het onderwijsaanbod. Voor de inhoud van de toetsen Spelling 3.0 zijn deze leerstoflijnen bepalend geweest, zowel als theoretische basis als voor de indeling van het categorieënoverzicht (zie verder hoofdstuk 3).

2.4.2 Psychometrisch

2.4.2.1 Opgavenbanken
Voor het samenstellen van toetsen voor het primair en speciaal onderwijs beschikt Cito over opgaven-banken. Die liggen ten grondslag aan onder meer de toetsen in het Cito Volgsysteem primair en speciaal onderwijs (LVS-toetsen). Voor de constructie van de toetsen Spelling 3.0 is gebruik gemaakt van de opgavenbank Spelling. Voor andere vakgebieden in het LVS zoals Begrijpend lezen, Woordenschat, Rekenen-Wiskunde en Begrijpend luisteren 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 spelling 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 spellingvaardigheid uit, waarbij een groter getal wijst op een grotere spellingvaardigheid. 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 spelling 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 zogenaamde p-waarde, de proportie correcte antwoorden op het item in een welbepaalde populatie van leerlingen. In de Item Respons Theorie (IRT) die voor het construeren van de opgavenbanken wordt gebruikt, hanteert men echter een andere definitie van moeilijkheid: ruwweg gesproken is het de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden. Dit verschil in definitie van de moeilijkheidsgraad tussen klassieke theorie en IRT is uitermate belangrijk: men kan verwachten dat de p-waarde van een item in groep 8 groter zal zijn dan in groep 6, waardoor duidelijk wordt dat de p-waarde een relatief begrip is: ze geeft de moeilijkheid aan van een item in een bepaalde populatie. Binnen de IRT is de moeilijkheid van een item gedefinieerd in termen van de onderliggende vaardigheid, zonder enige 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, waardoor het praktisch niet doenlijk 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 zogenaamde 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 zullen zijn. Voor een nauwkeuriger meting (bij een gegeven aantal items in de toets) moeten de moeilijkheidsgraden van de items in overeenstemming gebracht worden met het vaardigheidsniveau van de leerlingen.
2 Om een schatting te kunnen maken van de verdeling van de vaardigheid in een welomschreven populatie, worden selecties van items voorgelegd aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. In het geval van Spelling 3.0 zijn dat steekproeven van leerlingen op de verschillende normeringsmomenten vanaf medio groep 3 tot en met medio groep 8. Daarbij maakt het, behoudens wat bij 1 is vermeld over nauwkeurigheid, niet uit welke selectie van items aan een leerling binnen een normeringsgroep wordt afgenomen. Een van de eigenschappen van gekalibreerde itembanken is immers dat met elke selectie items de vaardigheid van leerlingen kan worden bepaald. Voor een voorbeeld hiervan, zie Staphorsius (1994). In de praktijk komt dit meestal neer op het schatten van gemiddelde en standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met deze schattingen kunnen dan ook schattingen gemaakt worden van de percentielen in de populatie.
3 Aan leerlingen die niet tot de betreffende referentiepopulatie behoren, kan dezelfde toets worden voorgelegd. De toetsscore wordt omgezet in een schatting van de vaardigheid en deze schatting kan geplaatst worden in de vaardigheidsverdeling van de populatie. Een leerling met achterstand in groep 4 kan een toets maken die normaliter aan groep 3 wordt voorgelegd, en zijn vaardigheidsschatting kan behalve met de populatie van groep 4 ook vergeleken worden met de percentielen in de populatie van groep 3, met bijvoorbeeld de uitspraak: “De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 3”.
4 De vergelijking die in het voorgaande gemaakt is, kan evengoed plaatsvinden als de (achterstands)-leerling een andere toets (i.e. een selectie uit de opgavenbank) maakt dan de toets die normaliter aan groep 3 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 Spelling. 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 zogenaamde ware score, de gemiddelde score die de persoon zou behalen indien de test een oneindig aantal keren onder dezelfde condities zou worden afgenomen. Die notie geeft een van de

belangrijkste (praktische) obstakels van deze theorie voor ons onderzoek weer: het is problematisch om toetsscores te vergelijken die verkregen zijn in een onvolledig design. Hoewel er methoden bestaan binnen de klassieke testtheorie om toetsscores te equivaleren (Engelen & Eggen, 1993), schiet deze benadering tekort als het gaat om de centrale vraag: hoe 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 *Xi* de toevalsvariabele die het antwoord op item *i* voorstelt. *Xi* neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord. Als symbool voor de vaardigheid wordt θ (theta) gekozen. De vaardigheid θ is niet rechtstreeks observeerbaar. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd3. De itemresponsfunctie *fi*(θ) 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 *fi*(θ) een meer of minder specifieke functionele vorm wordt toegekend. Een eenvoudig en zeer populair voorbeeld is het zogenaamde Raschmodel (Rasch, 1960) waarin *fi*(θ) gegeven is door

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

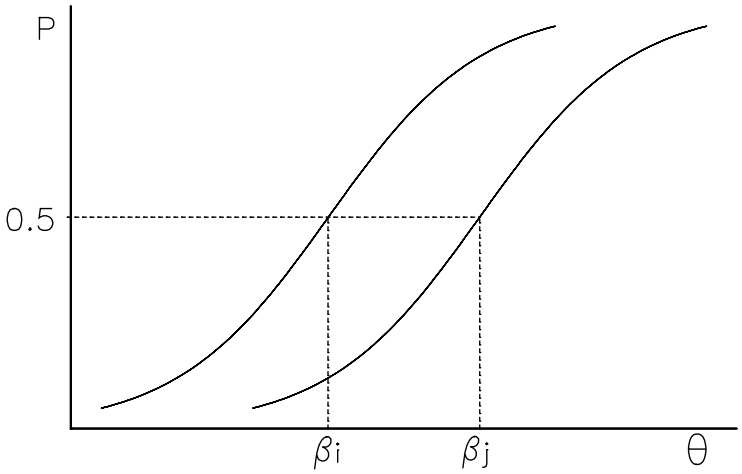
waarin β*i* de moeilijkheidsparameter van item *i* is. Dat is een onbekende grootheid die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.1 voor twee items, *i* en *j*, die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende functie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

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

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

[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 gebruik van deze functie. 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 zogeheten 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).

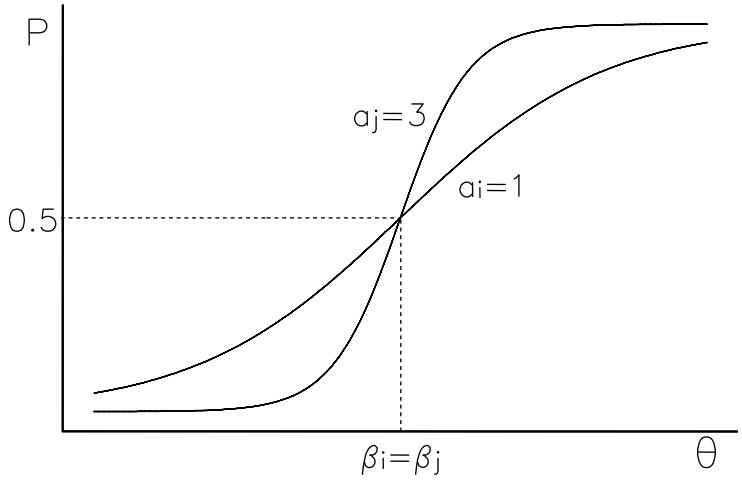
[4] Een gedetailleerde uiteenzetting hierover kan men vinden in Verhelst, 1992.

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

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



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

Voor de schatting van de populatieverdeling wordt gebruikgemaakt van de ‘marginale grootste aannemelijkheidsmethode’ (in het Engels: Marginal Maximum Likelihood, verder afgekort als MML). Deze schattingsmethode veronderstelt naast (2.2) ook nog dat de vaardigheid θ in de populatie een bepaalde verdeling heeft. De meeste computerprogramma’s die IRT-analyses kunnen uitvoeren, veronderstellen een normale verdeling. Bovendien stelt deze methode de voorwaarde dat de steekproef die voor de schatting gebruikt wordt uit die verdeling een aselecte steekproef is. Omdat leerlingen bovendien gevolgd worden, is het mogelijk gelijktijdig de verdelingen op de verschillende normeringsmomenten te schatten.

3 Beschrijving van de toetsen

3.1 Opbouw en structuur van de toetsen

Het toetspakket Spelling 3.0 voor groep 3 uit het Cito Volsysteem primair en speciaal onderwijs bevat in totaal drie papieren toetsen: M3, M3E3 en E3. De toetsen M3 en E3 zijn de reguliere toetsen, bedoeld voor afname op de reguliere afnamemomenten medio (half januari/half februari) of einde (juni) schooljaar. Naast deze reguliere toetsen bevat het toetspakket ook een extra toets, M3E3. Ook deze extra toets is bedoeld om af te nemen op de reguliere afnamemomenten. Door het toevoegen van een extra toets (tussentoets) beslaan de opeenvolgende toetsen kleinere leerstappen. Daardoor zijn er voor speciale leerlingen meer toetsen beschikbaar voor het meten van de spellingvaardigheid. Qua moeilijkheid zit deze extra toets tussen de toetsen M3 en E3 in. De extra toets kan voorgelegd worden aan leerlingen voor wie de toets E3 op het afnamemoment einde schooljaar net te moeilijk is.

Opbouw
De toetsen Spelling 3.0 voor groep 3 bestaan telkens uit twee taken van elk 20 opgaven. Deze dienen bij voorkeur te worden afgenomen op twee verschillende dagdelen, zodat de leerlingen geconcentreerd aan beide taken kunnen werken.

Vorm
De taken bevatten in groep 3 twee soorten opgaven:
– woorddictee (M3);
– zinsdictee (M3E3 en E3).

De eerste toets Spelling, te weten M3 (bedoeld voor afname in januari van groep 3), bevat alleen woorddicteeopgaven. Bij het woorddictee leest de leerkracht losse woorden voor. De leerlingen schrijven die woorden in een opgavenboekje. In het opgavenboekje staat bij elk opgavennummer een illustratie die past bij het voorgelezen woord. De illustraties zijn bedoeld om mogelijke twijfel bij de leerlingen weg te nemen over welk woord de leerkracht nu precies zei. Omdat de woorden niet in een context worden aangeboden, zou een onduidelijke uitspraak van de leerkracht ertoe kunnen leiden dat de leerling bijvoorbeeld 'rook' schrijft in plaats van 'rok'. Bij de keuze voor de illustraties is rekening gehouden met de mate waarin leerlingen van groep 3 abstract kunnen denken. Ook zijn de behoeften van speciale leerlingen in aanmerking genomen. De tekeningen zijn daarom zo concreet en helder mogelijk en bevatten zo min mogelijk details.
Een woorddictee halverwege groep 3 sluit aan bij het aanbod van de leerstof in de methoden. Goed luisteren naar het dicteewoord vergt al alle aandacht van de leerling in de aanvangsfase van het leren lezen en spellen. Een woord in een zin aanbieden kan dan te veel afleiden. Om de leerlingen te laten focussen op het woord, krijgen ze elk woord afzonderlijk aangeboden in een woorddictee.

De toetsen vanaf eind groep 3 bevatten alleen zinsdicteeopgaven. Bij een zinsdictee leest de leerkracht een zin voor en herhaalt vervolgens uit deze zin één woord. Dat woord moeten de leerlingen opschrijven. Door de dicteewoorden in zinsverband aan te bieden is twijfel over het bedoelde dicteewoord vrijwel uitgesloten en kunnen illustraties achterwege blijven.

In het toetsen van spellingvaardigheid is onderscheid te maken tussen actieve spelling (dicteeopgaven) en passieve spelling (meerkeuzeopgaven). In de derde generatie toetsen Spelling van het Cito Volgsysteem primair en speciaal onderwijs is ervoor gekozen om alleen actieve spelling op te nemen. Dat wil zeggen dat er dus geen meerkeuzeopgaven zijn opgenomen zoals in de toetsen Spelling van de tweede generatie van het Cito Volgsysteem nog wel het geval was. Door alleen dicteeopgaven op te nemen, is gehoor gegeven aan een wens vanuit het veld. Leerkrachten hadden met name voor de onderbouw liever geen meerkeuze-opgaven, omdat leerlingen bij deze opgaven geconfronteerd worden met foutieve woordbeelden.

Het omschakelen naar uitsluitend dicteeopgaven levert geen problemen op voor de meetpretentie van de toetsen. Analyses op de itemgegevens van de tweede generatie toetsen lieten zien dat dicteeopgaven en meerkeuzeopgaven op één schaal pasten. We toetsen nog steeds dezelfde vaardigheid: er is continuïteit tussen de toetsen uit de tweede generatie en de nieuwe toetsen, zoals ook blijkt uit de resultaten van analyses die we in hoofdstuk 6 presenteren.
In de bovenbouw van het basisonderwijs worden leerlingen steeds meer geacht hun eigen schrijfwerk – en vaak ook dat van medeleerlingen – na te kijken. Vandaar dat passieve spelling hier ook van belang is. Daarom is passieve spelling, naast grammatica en leestekens, ondergebracht in de toetsen Taalverzorging voor groep 6 tot en met 8 (beschikbaar vanaf schooljaar 2015/2016). Doordat actieve en passieve spelling in de toetsen voor de bovenbouw uit elkaar worden gehaald, kan op beide onderdelen apart gerapporteerd worden.

Keuze van een passende toets: toetsen op maat
De vaardigheid in spelling van leerlingen in een groep loopt vaak sterk uiteen. Als gevolg daarvan zal eenzelfde toets spelling voor een deel van de leerlingen op niveau zijn, maar voor sommige 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 E3-toets voor leerlingen eind groep 3) 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 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 Spelling deze score ook behaald is – kan worden omgezet in een score op één en dezelfde vaardigheidsschaal. Leerlingen kunnen daardoor bijvoorbeeld een toets maken die hoort bij een vorig afnamemoment (een M4-leerling maakt een toets E3) of een volgend afnamemoment (een E3-leerling maakt de toets M4).
Voor leerlingen met een vertraagde ontwikkeling is de extra toets M3E3 in het toetspakket voor groep 3 opgenomen.

Afname
De toets wordt in principe klassikaal afgenomen door de leerkracht of IB'er. De afname start met een klassikale instructie door de leerkracht of IB'er. De leerkrachtmap bevat hiervoor een afnamekaart met afname-instructies aan de hand van een voorbeeldopgave. De leerkracht of IB'er leest vervolgens één voor één de woorden of zinnen van de toets voor en de leerlingen schrijven het woord op dat de leerkracht/IB'er nog een keer herhaalt. De afname is niet aan tijd gebonden. Als een leerling het dicteewoord niet goed gehoord heeft, mag de leerkracht/IB'er de opgave nog een keer herhalen. Aan het eind van het dictee kijken de leerlingen hun opgaven nog een keer na. Daarna haalt de leerkracht/IB'er de antwoordboekjes of -bladen op.
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 leerkrachtmap 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. Er is geëxpliciteerd 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 ook gebruikgemaakt worden van de afnamekaarten. Hierop staan gedetailleerde scoringsvoorschriften en nakijkinstructies. Indien gewenst kan de leerkracht in het Computerprogramma de foute antwoorden invoeren of aanklikken. Op basis van het aantal goede antwoorden, de toetsscore, wordt een inschatting gemaakt van de vaardigheid van de leerlingen.
De leerkracht 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 leerkrachtmap 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, zoals leerlingrapporten en groepsoverzichten. Deze rapportages zijn beschikbaar via Cito Portal.
Daarnaast is er via Cito Portal voor elke toets een analyseformulier Spelling beschikbaar. Als de leerkracht de fouten van leerlingen verder wil analyseren, kan hij gebruik maken van dit formulier.

Niet alleen de scoring van de toetsen kan met behulp van de computer worden uitgevoerd, ook de foutenanalyse kan via de computer gedaan worden. Dit kan door de fout gespelde woorden in te voeren in het Computerprogramma LOVS. In het computerprogramma vindt een automatische analyse plaats van de gemaakte fouten. Deze worden direct bij de juiste spellingcategorie(ën) ingedeeld. Net als bij de papieren foutenanalyse wordt hier aangegeven of het om de beoogde categoriefout gaat of om een andere categorie en om welke categorie het dan gaat. Als de leerling een fout heeft gemaakt die niet in een categorie in te delen is, wordt deze fout aangegeven als 'andere fout'. In groep 3 is het mogelijk naast een analyse van de spellingcategorieën ook een analyse te doen van het aanvankelijk schrijven. Ook hiervoor wordt de analyse in het computerprogramma automatisch uitgevoerd.

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 leerkrachtmap 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 mogelijkheid om functioneringsniveaus op te vragen. De functioneringsniveaus geven aan met welke gemiddelde leerling de vaardigheidsscore van de getoetste leerling vergelijkbaar is. Een functioneringsniveau E3 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling overeenkomt met de score van de gemiddelde leerling eind groep 3. 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:
- domeinbeschrijving en uitwerking in spellingcategorieën;
- 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 is opgenomen in de handleiding van het toetspakket Spelling 3.0 voor groep 3 (hoofdstuk 6). Daarin staat een uitgebreide beschrijving van de methode-analyse en een overzicht per jaargroep en afnamemoment van alle getoetste woorden. Ook informatie over de moeilijkheid van elk getoetst woord is daar te vinden (in de vorm van grafieken).

3.2.1 Domeinbeschrijving en uitwerking in spellingcategorieën

Omdat het Nederlandse spellingsysteem gebaseerd is op verschillende principes, hebben kinderen een hele weg te gaan om goed te leren spellen. Bij het (leren) spellen kunnen vaak verschillende strategieën worden ingezet, zoals bijvoorbeeld de fonologische strategie (opschrijven wat je hoort). Wat wij met de toetsen Spelling 3.0 beogen te meten is of leerlingen weten hoe een woord correct gespeld moet worden. Hoe leerlingen daarbij te werk gaan, is voor dit doel niet interessant. Er leiden immers verschillende wegen naar Rome …

Bepaalde woorden zijn eenvoudiger correct te spellen dan andere woorden. Dit wordt ook in het onderwijs onderkend: alle spellingmethoden kennen een opbouw van gemakkelijker te spellen woorden naar moeilijker te spellen woorden. De criteria voor het ordenen van de leerstof staan genoemd in hoofdstuk 2. Ook bij het toetsen van de vaardigheid spelling gaan we uit van een indeling van gemakkelijk (eenlettergrepig, klankzuiver) naar moeilijk (meerlettergrepig, niet klankzuiver).

In hoofdstuk 2 hebben wij de theoretische uitgangspunten van de toetsen Spelling beschreven. We hebben daarbij gebruikgemaakt van wetenschappelijke publicaties over spelling en het wettelijke referentiekader Nederlandse taal (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a), inclusief de Leerstoflijnen Begrippenlijst en taalverzorging (Van der Beek & Paus, 2011).

De taalkundige indeling in vijf klassen van spellingproblemen heeft de basis gevormd voor een overzicht van spellingcategorieën. Hiermee sluiten de toetsen Spelling 3.0 aan bij de indeling die is gehanteerd in het Referentiekader Taal bij het domein ‘Begrippenlijst en Taalverzorging’ (Expertgroep Doorlopende Leerlijnen taal en rekenen, 2009; Van der Beek & Paus, 2011). Op basis van een methode-analyse hebben we de indeling van spellingcategorieën vervolgens verder verfijnd tot een in- en verdeling van spellingcategorieën over alle toetsen Spelling 3.0 voor groep 3 tot en met 8.

Hieronder wordt de totstandkoming en onderbouwing van dit categorieënoverzicht nader toegelicht.

Methodeanalyse
Het overzicht van de spellingcategorieën voor groep 3 tot en met 8 is ontwikkeld op basis van een uitgebreide methodeanalyse. Voor deze analyse hebben we in 2010 en 2013/2014 de spellingleergangen van acht verschillende taalmethoden met elkaar vergeleken. Verder hebben we verschillende methoden voor aanvankelijk lezen bekeken. In tabel 3.1 staan de methoden die in de analyse zijn meegenomen.

Tabel 3.1 Onderzochte methoden

Methode	Uitgever	Jaar van uitgave
De Leessleutel	Uitgeverij Malmberg, 's-Hertogenbosch	2002
Lijn 3	Uitgeverij Malmberg, 's-Hertogenbosch	2014
Spelling in beeld	Uitgeverij Zwijsen B.V., Tilburg	2006, 2013
Spelling op maat	Noordhoff, Houten	2006, 2013
Staal	Uitgeverij Malmberg, 's-Hertogenbosch	2013
Taal actief	Uitgeverij Malmberg, 's-Hertogenbosch	2003, 2012
Taaljournaal	Uitgeverij Malmberg, 's-Hertogenbosch	2005
Taalleesland	Bekadidact, Baarn	2004-2007
Taalverhaal	ThiemeMeulenhoff, Utrecht	2002-2003, 2013
Veilig leren lezen	Uitgeverij Zwijsen B.V., Tilburg	2003
Zin in spelling	Uitgeverij Zwijsen B.V., Tilburg	2006

Bij een aantal methoden staat het jaar van uitgave van twee edities. In onze methode-analyse van 2010 hebben we de eerdere edities bekeken, in 2013/2014 hebben we in een aanvullende methode-analyse de nieuwste edities bekeken. De definitieve gegevens van de methodeanalyse zijn gebaseerd op beide edities.

In 2013 heeft Cito een rapport gepubliceerd in het kader van de Periodieke Peiling van het Onderwijsniveau op het gebied van schrijfvaardigheid in het basisonderwijs (Kuhlemeier, Van Til, Hemker, De Klijn & Feenstra, 2013). Uit dit rapport blijkt dat bovengenoemde taalmethoden in het basisonderwijs het meest gebruikt worden. In dit rapport was 'Staal' nog niet opgenomen. Deze methode is namelijk in 2013 voor het eerst verschenen. In ons aanvullend onderzoek van 2013/2014 hebben we deze methode wel meegenomen. Ook de aanvankelijk leesmethoden waren niet in het rapport opgenomen. Alle onderzochte methoden tezamen zorgen voor een zeer hoge dekkingsgraad.

Alle woorden in de spellingtoetsen horen bij een bepaalde spellingcategorie. Een spellingcategorie geeft aan welke spellingmoeilijkheid er in het woord zit. Bij de analyse zijn we uitgegaan van de referentiekaders voor Spelling. Hiervoor hebben we de rapporten van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008a, 2008b, 2009a, 2009b) en de uitgave 'Leerstoflijnen begrippenlijst en taalverzorging beschreven' van SLO (Van der Beek & Paus, 2011) bestudeerd. Het categorieënoverzicht LOVS Spelling (2006-2010) gebruikten we als uitgangspunt voor het nieuwe categorieënoverzicht.

Om vast te stellen in welke toets een spellingcategorie voor het eerst zou moeten voorkomen, is nagegaan in welk jaar en op welk moment een spellingkwestie in de meeste methoden behandeld was.
Dit analyseerden we eerst per methode; vervolgens vergeleken we de methoden met elkaar. We geven hier een compleet overzicht van de resultaten van de methodeanalyse voor groep 3 en groep 4 samen voor het volledige beeld.

Voor groep 3 konden we ons beperken tot de aanvankelijk leesmethoden, spelling komt hierin ook aan bod. Leren lezen hangt immers nauw samen met leren spellen. De leerlingen leren in groep 3 de letters. Aan de hand van de foneem-grafeemkoppeling leren ze woorden te herkennen en zelf te schrijven. De aanvankelijk leesmethoden richten zich op lezen, maar maken ook ruimte in de methode voor het zelf schrijven van woorden. Van deze methoden heeft *Veilig leren lezen* de hoogste dekkingsgraad in groep 3. Na *Veilig leren lezen* wordt *De Leessleutel* het meest gebruikt in het onderwijs. *Lijn 3* is de opvolger van *De Leessleutel*. Deze methode is in 2014 uitgekomen.

Vanaf groep 4 zijn er methoden die zich specifiek op spelling richten. Als een spellingcategorie (vanaf groep 4) in vijf van de acht methoden aan bod was geweest, werd de categorie opgenomen in de toets. De consequentie hiervan is dat er een enkele keer categorieën in de toetsen zijn opgenomen die op het moment van toetsafname nog niet in alle methoden aan bod gekomen zijn. Toch hebben wij er bewust voor

gekozen niet te wachten met het opnemen van een categorie tot deze in alle acht onderzochte methoden behandeld zou zijn. Een ongewenst gevolg daarvan zou namelijk kunnen zijn dat veel categorieën pas in de toets aan bod komen op het moment dat (bijna) alle leerlingen ze volledig beheersen. Het is dan minder goed mogelijk om de precieze vaardigheid van een leerling vast te stellen. De toets is dan een beheersingstoets geworden. Een andere ongewenste consequentie zou kunnen zijn dat er op het ene moment te veel categorieën in de toets aan bod komen en op een ander moment heel weinig. Voor de aansluiting van de toetsen Spelling 3.0 bij het gegeven onderwijs zou het natuurlijk ideaal zijn als alle taalmethoden eenzelfde aanbiedingsvolgorde van spellingcategorieën zouden hanteren. Maar dat is nu eenmaal niet het geval.
In de handleiding bij de toetsen Spelling 3.0 is aangegeven dat leerkrachten hun onderwijsaanbod niet hoeven aan te passen op het moment dat een spellingcategorie nog niet aan bod is geweest. Omdat de resultaten wel iets kunnen afwijken, wordt het advies gegeven wel een aantekening te maken van dit verschil in onderwijsaanbod. Als de spellingcategorie eenmaal aan bod is geweest, herstelt het resultaat zich op een volgend toetsmoment vanzelf weer.

Spellingcategorieën in de toetsen Spelling 3.0 groep 3
Alle spellingcategorieën zijn uiteindelijk bij elkaar gezet in een categorieënoverzicht. In het toetspakket is het categorieënoverzicht voor groep 3 en 4 opgenomen, als bijlage bij de handleiding. Hierin is aangegeven op welk(e) moment(en) een bepaalde categorie aan bod komt in de toetsen Spelling. Als categorieën op meerdere toetsmomenten aan bod komen, is dit duidelijk zichtbaar.
Per toets komt slechts een deel van het totaal aantal spellingcategorieën aan de orde. In tabel 3.2 is te zien welke categorieën deel uitmaken van de toetsen Spelling voor groep 3 en hoeveel opgaven het per toets betreft. Er komen negen spellingcategorieën in het onderwijs en in de toetsen aan bod. De opgaven zijn verdeeld over de twee taken waaruit elke toets bestaat. In de handleiding staan de precieze aantallen en woorden per categorie en per taak.

Tabel 3.2 Spellingcategorieën in de toetsen Spelling 3.0 groep 3

Cat.	**Omschrijving**	**M3**	**M3E3**	**E3**
1	mkm-woorden	14	3	
2	mmkm- en mkmm-woorden	26	6	4
3	mmkmm-woorden		7	7
4	woorden met een niet geschreven tussenklank		7	5
5	woorden met -mmm of mmm-		6	7
6	woorden met sch(r)-		5	5
7	woorden met -ng(-) of -nk(-)		6	5
8	woorden met f-, v-, s- of z-			3
9	verkleinwoord met uitgang -je(s) of -tje(s)			4

De totale aantallen opgaven per toets zijn gemakkelijk uit bovenstaande tabel af te leiden door de aantallen opgaven per kolom bij elkaar op te tellen. De toetsen in groep 3 bevatten elk 40 opgaven, 20 opgaven per taak.

3.2.2 Itemconstructie, onderzoeken en selectie van opgaven voor de toetsen Spelling

Itemconstructie
Alle opgaven die in de toetsen Spelling zijn opgenomen werden voor deze toetsen geconstrueerd door een speciaal hiervoor samengestelde constructiegroep. De groep bestond uit leerkrachten uit het basisonderwijs en een spellingexpert. De constructiegroepleden selecteerden woorden bij de verschillende spellingcategorieën. Om zeker te weten dat een geselecteerd woord bekend is bij leerlingen van groep 3, is telkens nagegaan of het woord voorkwam in Digiwak. Digiwak is een digitale woordenlijst. Deze bevat alle

woorden die leerlingen aangeboden hebben gekregen in het onderwijs. De woorden zijn ingedeeld naar thema en naar groep (gebaseerd op Kuiken & Droge (2010)). Op enkele uitzonderingen na bevatten de toetsen Spelling woorden waar de leerlingen bekend mee zijn.
Een toetsdeskundige verzorgde in samenwerking met een illustrator de tekeningen bij de woorden voor M3. De constructiegroepleden maakten dicteezinnen bij de woorden vanaf M3E3. Zij kregen de instructie de dicteezinnen zo kort en concreet mogelijk te maken en de plaatjes eenvoudig te houden, zodat tegemoet gekomen wordt aan leerlingen met speciale leerbehoeften. Het gaat hier om leerlingen met een vertraagde ontwikkeling, een beperkte aandachtsspanne of een grote behoefte aan structuur. In de dicteezinnen zijn de ik-vorm, gebiedende wijs en vraagzinnen zoveel mogelijk vermeden. Sommige leerlingen kunnen namelijk afgeleid worden door een vraagzin of gebiedende wijs. Bijvoorbeeld in zinnetjes als: 'Wie is er aan de beurt?' De leerling heeft dan de neiging om antwoord te geven op de vraag. Of: 'Ga eens rechtop zitten!'. Ook in dat geval kan de leerling zich aangesproken voelen. Zinnen in de ik-vorm zijn ook niet opgenomen, omdat met name leerlingen met een stoornis in het autistische spectrum deze zinnen snel op zichzelf betrekken.

Proeftoetsing en kalibratie-analyses
De opgaven zijn eerst in proefafnames voorgelegd aan leerlingen in de jaargroep 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 spellers dan door minder goede spellers fout gemaakt worden) geïdentificeerd en verwijderd worden. Daarnaast hebben wij de proefafname aangegrepen als een mogelijkheid om aan de deelnemende leerkrachten te vragen of zij inhoudelijke of andersoortige bezwaren hadden tegen bepaalde opgaven of dicteewoorden.
Bij proeftoetsing zijn in 2012 halverwege en aan het einde van leerjaar 3 een aantal opgaven voorgelegd aan leerlingen. Daarbij zijn op het afnamemoment medio groep 3 190 nieuwe opgaven geproeftoetst en op het afnamemoment eind groep 3 225. Elke deelnemende school maakte één taak met 25 nieuwe opgaven, naast de LVS-toets Spelling van de tweede generatie. In totaal waren de nieuwe opgaven op beide afnamemomenten verdeeld over 12 boekjes. Elk boekje werd door ongeveer 150 leerlingen gemaakt. Door de nieuwe opgaven in één of twee boekjes op te nemen, werd ervoor gezorgd dat elke opgave door minimaal 200 leerlingen werd gemaakt.
Na de afnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programmapakket One Parameter Logistic Model (OPLM; Verhelst, 1993; Verhelst en Glas, 1995). Zie voor een algemene technische beschrijving van dit model paragraaf 2.4.2.
Bij de analyses is de kwaliteit van de afzonderlijke items en van de totale opgavenverzameling 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 gescreend door praktijkdeskundigen uit het SBO. Hierbij is erop gelet dat de opgaven geschikt zijn voor een zo groot mogelijke groep leerlingen, ook voor leerlingen met extra onderwijsbehoeften. Opgaven waar de praktijkdeskundigen opmerkingen bij hadden, hebben we waar mogelijk aangepast of verwijderd.

Normeringsonderzoek
Op basis van de psychometrische analyses en de evaluaties van de proeftoetsing hebben we de opgaven geselecteerd voor het normeringsonderzoek. De psychometrische criteria betroffen met name de moeilijkheidsgraad en discriminatieparameter. Voor een evenwichtige samenstelling van de toetsen hebben we gelet op de verdeling van items over de verschillende spellingcategorieën.
Waar mogelijk hebben we bij de opgavenselectie rekening gehouden met de opmerkingen die de leerkrachten gemaakt hebben. Er waren soms woorden waarvan meerdere leerkrachten aangaven ze niet geschikt te vinden voor leerlingen in groep 3. Meestal is dat woord dan ook niet opgenomen in het normeringsonderzoek. In een enkel geval zijn we daarvan afgeweken, omdat er anders te weinig geschikte woorden zouden overblijven in de betreffende categorie. Verder gaven sommige leerkrachten aan dat zij

bepaalde opgaven nogal moeilijk vonden. Dit is een bekend fenomeen bij toetsen voor (zeer) jonge kinderen. Bij de LVS-toetsen Spelling van de tweede generatie bleek echter al dat de meeste leerlingen in groep 3 meer woorden goed spellen dan verwacht. Uit de proeftoetsgegevens van Spelling 3.0 kwam dit eveneens naar voren. In die gevallen dat de genoemde opgaven een goede p-waarde hadden, zijn die dicteewoorden toch opgenomen. Als ze daadwerkelijk te moeilijk waren, zijn ze verwijderd uit de toets. Alle opgaven met een acceptabele moeilijkheid (in klassieke termen een p-waarde tussen .40 en .90) die door de betere spellers significant vaker goed werden gemaakt dan door de minder goede spellers (rir vanaf .20) kwamen in principe in aanmerking voor opname in de normeringsonderzoeken Spelling. Voor sommige spellingcategorieën bleken er na afloop van de proefafname te weinig psychometrisch acceptabele opgaven overgebleven te zijn. Voor die categorieën werden na de proefafname nog nieuwe opgaven geconstrueerd.

De opgaven die na de proefafname geselecteerd waren plus een aantal extra geconstrueerde opgaven werden vervolgens ingedeeld voor opname in de normeringsonderzoeken.
In tegenstelling tot de proefafnames, waar opgaven random over toetsboekjes werden verdeeld, zijn in de normeringsonderzoeken de taken zodanig samengesteld dat ze al zoveel mogelijk leken op de definitief uit te geven taken. In elke taak zaten opgaven van uiteenlopende moeilijkheid. Elke taak bevatte opgaven uit alle te toetsen spellingcategorieën, in een evenwichtige verdeling. In de taken zijn er bijvoorbeeld geen opgaven van dezelfde categorie direct na elkaar geplaatst. Ook woorden met een zelfde beginletter of beginklank zijn verdeeld over de taken. Verder is ervoor gezorgd dat de taken begonnen en eindigden met gemakkelijke opgaven. Ondanks zorgvuldige proeftoetsing van opgaven kan het voorkomen dat sommige dicteewoorden in een normeringsonderzoek alsnog onvoldoende functioneren. Om die reden bevatte elke taak enkele opgaven méér dan de definitieve taken, met het oog op eventuele uitval van opgaven.

Samenstelling definitieve toetsen
Na het normeringsonderzoek is van alle opgaven opnieuw de p-waarde en de rir bepaald. Ook nu kwamen in principe alle opgaven met een acceptabele moeilijkheid (in klassieke termen een p-waarde tussen .40 en .90) die door de betere spellers significant vaker goed werden gemaakt dan door de minder goede spellers (rir vanaf .20) in aanmerking voor opname in de definitieve toetsen Spelling. Verder is opnieuw gekeken naar de indeling in spellingcategorieën. Vanaf de toets M3E3 komen in sommige woorden meer dan één spellingcategorie aan de orde. In alle gevallen is het woord vóór de proefafname ondergebracht bij de categorie die het laatst aan de orde is geweest in het spellingonderwijs. Om de definitieve categorie van een woord vast te stellen, hebben we geïnventariseerd hoe vaak een bepaalde spellingfout daadwerkelijk gemaakt werd door leerlingen. Hiervoor gebruikten we frequentielijsten: lijsten waarin alle verschillende gegeven antwoorden, voorzien van frequentie van vóórkomen, opgesomd zijn. In een enkel geval bleek een fout van een (iets) lagere categorie of een fout in het aanvankelijk schrijven vaker gemaakt te worden dan de beoogde fout van de hogere categorie. In het woord ‘school’ bijvoorbeeld was categorie 6 (woorden met sch(r)-) de beoogde categorie. Veel leerlingen maakten in dit woord echter een fout in het aanvankelijk schrijven: klinkerverenkeling (ze schreven ‘schol’ in plaats van ‘school’). Dit woord is uiteindelijk wel opgenomen in de toets, omdat ook de categoriefout erg vaak gemaakt werd. Voor de woorden die we opgenomen hebben in de toetsen hebben we in de meeste gevallen kunnen vaststellen dat er in het normeringsonderzoek vooral fouten gemaakt werden in de beoogde categorie. Het is natuurlijk nog steeds mogelijk dat een leerling behalve de categoriefout ook een fout maakt in een andere spellingcategorie.
Op Cito-Portal hebben we analyseformulieren Spelling opgenomen. De leerkracht heeft de mogelijkheid met behulp van deze formulieren een diepgaandere analyse uit te voeren.

Voor groep 3 was het lastig woorden te vinden die moeilijk genoeg waren voor de definitieve toetsen. Veel woorden in de proeftoets hadden een p-waarde van boven de 0,90. Voor het normeringsonderzoek waren de beste opgaven geselecteerd, maar nog steeds waren veel opgaven erg gemakkelijk voor de leerlingen van groep 3. De meeste leerlingen beheersen op dat moment al de spellingcategorieën die aangeboden worden in de toetsen. Een optie om de toetsen op een acceptabel moeilijkheidsniveau te krijgen was het opnemen van extra categorieën. Een andere optie was het inkorten van de toetsen zodat de makkelijkste items konden vervallen.

We hebben ervoor gekozen géén extra categorieën aan te bieden in de toetsen, omdat we dan vooruit zouden lopen op het spellingonderwijs. Het voordeel daarentegen van de tweede optie – het inkorten van de toetsen – was dat kortere toetsen beter passen bij jonge leerlingen, aangezien hun spanningsboog nog niet zo groot is. Oorspronkelijk was het de bedoeling om toetsen van 50 opgaven (25 opgaven per taak) te maken. De uiteindelijke toetsen voor groep 3 bevatten 40 opgaven (20 opgaven per taak). Door de toetsen in te korten was het mogelijk om de beste opgaven te selecteren en een deel van de opgaven met een p-waarde van 0,90 of meer te verwijderen. De hoge betrouwbaarheid van de toetsen kon op deze manier behouden blijven (zie verder hoofdstuk 5). Omdat de toetsen opgaven bevatten van uiteenlopende moeilijkheid, zijn ze geschikt om verschillen tussen leerlingen in beeld te brengen en worden de echte probleemspellers met deze toetsen gesignaleerd. Bij het bepalen van de volgorde van de opgaven in de definitieve toets, hebben we zo min mogelijk afgeweken van de volgorde in het normeringsonderzoek. Uit extra analyses bleek overigens dat er geen volgorde-effecten optraden.
Met de toetsen van groep 3 sluiten we dus aan bij het onderwijsaanbod in groep 3. De toetsen passen optimaal bij de jonge doelgroep: ze zijn kort en ze maken de toetservaring voor deze jonge kinderen die aan het begin van hun onderwijscarrière staan tot een positieve ervaring.

Bij het samenstellen van de definitieve toetsen zijn de volgende *inhoudelijke* criteria aangehouden:
1. Als in de spellingmethoden in een bepaald leerjaar bepaalde spellingcategorieën werden behandeld, dan wilden wij die categorieën op het eerstvolgende afnamemoment in de toets terug laten komen.
2. Het aantal categorieën dat op enig afnamemoment in een toets Spelling aan de orde kwam, mocht niet zodanig hoog zijn dat de leerling minder dan drie opgaven per categorie kreeg voorgelegd.
3. De verdeling van opgaven over categorieën en taken moest zo gelijkmatig mogelijk zijn.

De eerste twee criteria leverden geen problemen op. Het derde criterium, een zo gelijkmatig mogelijke verdeling van het aantal opgaven per categorie, is overal waar mogelijk aangehouden. Er is telkens opnieuw een afweging gemaakt op basis van inhoudelijke en psychometrische informatie, waarbij psychometrische argumenten het soms wonnen van inhoudelijke. In een paar gevallen is daarom de verdeling van opgaven niet helemaal gelijkmatig.
De uiteindelijke verdeling van aantallen opgaven per categorie per afnamemoment is een zo goed mogelijk compromis. Hierbij willen we graag benadrukken dat alle items, ongeacht de spellingcategorie waarop ze betrekking hebben, goed passen op de onderliggende vaardigheidsschaal.

Een illustratie van de samenstelling en moeilijkheidsgraad van de toetsen zijn de figuren in bijlage 2: p50 en p80-kanspunten van de opgaven in de toetsen voor groep 3 in relatie tot de gemiddelde vaardigheidsscore voor de afnamemomenten. 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. In deze figuren is zichtbaar dat de toetsen opgaven bevatten van uiteenlopende moeilijkheidsgraad.
Bij de toets M3 zijn er veel makkelijke opgaven (die liggen onder de stippellijn van M3) en een paar moeilijkere opgaven (doorkruisen de lijn van M3 of liggen erboven). De toets is weliswaar makkelijk, maar de echte probleemspellers worden wel gesignaleerd.
Ook is te zien dat de toets E3 relatief iets moeilijker is dan de toets M3: er liggen bij de toets E3 relatief meer balkjes op en rond de gemiddelde vaardigheidsscore.
Bij de gemakkelijke variant van de toets E3, de toets M3E3, liggen meer balkjes ónder de lijn van E3 dan bij de toets E3 (dus onder de gemiddelde vaardigheidsscore). Deze toets is over de gehele linie erg makkelijk en dus geschikt voor de zwakkere spellers.

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 het geval te zijn. Voor uitgebreide informatie over de kalibratie verwijzen wij naar hoofdstuk 4.

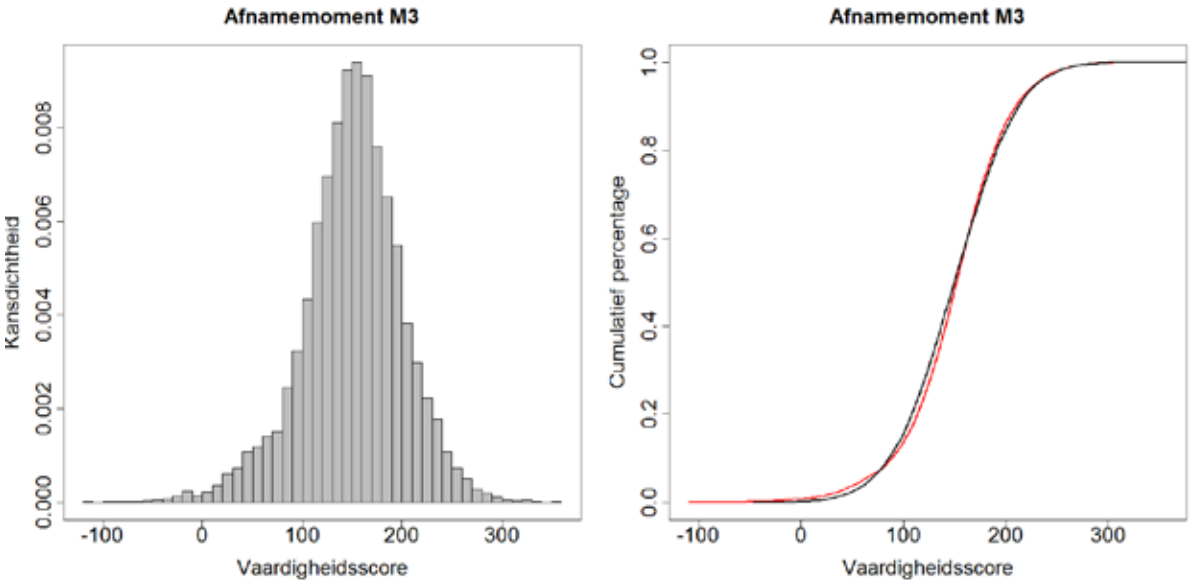
3.3 Statistische beschrijving

In hoofdstuk 4 zullen de kalibratie en normering uitgebreid worden beschreven. We geven hier alvast een samenvattend overzicht van de beschrijvende gegevens van de toetsen voor groep 3, zowel op de ruwe scoreschaal als op de vaardigheidsschaal. In tabel 3.3 valt op dat de vaardigheidsverdeling van de extra toets M3E3 gelijk is aan die van E3. De reden is dat voor deze extra toets de normeringsgegevens van de genormeerde toets E3 zijn gebruikt. De gegevens zijn gebaseerd op 2964 leerlingen voor M3 en op 2996 leerlingen voor E3. De waarden laten zien dat de vaardigheidsverdeling bij benadering normaal is. Figuur 3.1 met de verdeling van de vaardigheidsscores van M3 en figuur 3.2 met de verdeling van de vaardigheidsscores van E3 laten dit ook zien.

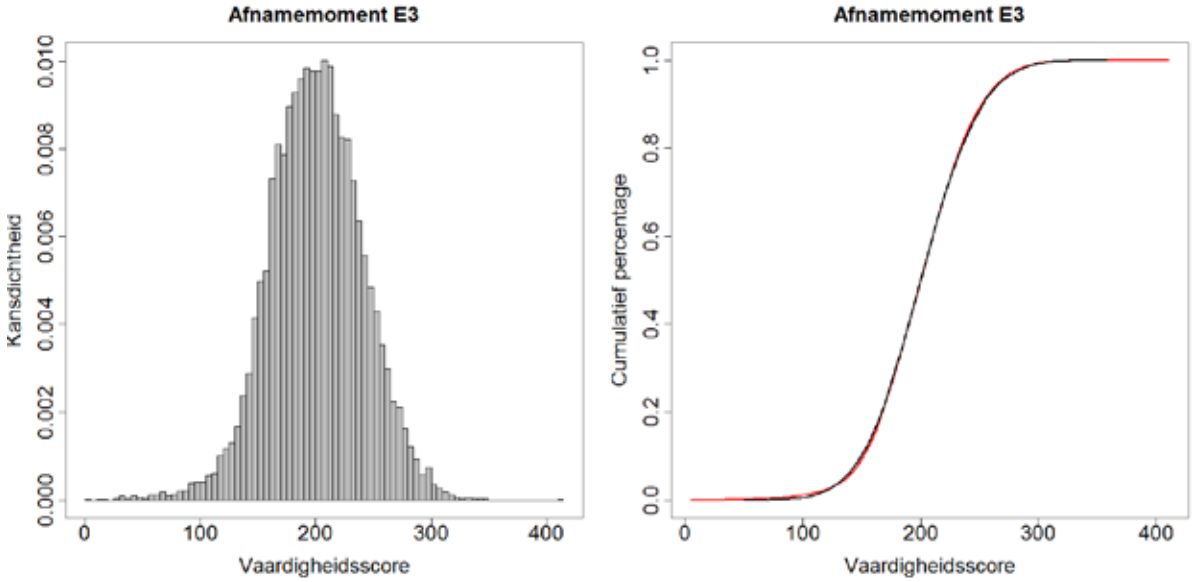
Tabel 3.3 Beschrijvende gegevens toets M3, M3E3 en E3 op ruwe scoreschaal en vaardigheidsschaal

	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
M3 Ruwe score	31,4	7,7	0,76	-1,17
M3 Vaardigheid	150,0	50,0	1,05	-0,38
M3E3 Ruwe score	34,8	5,1	3,35	-1,67
M3E3 Vaardigheid	200,4	40,4	0,63	-0,14
E3 Ruwe score	32,0	6,1	1,23	-1,16
E3 Vaardigheid	200,4	40,4	0,63	-0,14

Figuur 3.1 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling M3



Figuur 3.2 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling E3



4 Kalibratie en normering

4.1 Opzet voor de normeringsonderzoeken van het LVS: het macro-design

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 (M3, E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8) over normen te beschikken, er moet ook voor gezorgd worden dat de prestaties over de jaren heen met elkaar vergelijkbaar zijn. Hiertoe dienen de prestaties van leerlingen over alle leerjaren heen te worden afgebeeld op een gemeenschappelijke vaardigheidsschaal. Om zo'n gemeenschappelijke schaal te realiseren kunnen we niet volstaan met het ontwikkelen van afzonderlijke toetsen voor de meetmomenten en elke toets afzonderlijk ijken en normeren. Prestaties van bijvoorbeeld de populatie M5 moeten vergelijkbaar zijn met die van andere populaties, bijvoorbeeld E4 en E5, oftewel het 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.
Het macro-design is weergegeven in figuur 4.1.

		M3		E3		M4		E4		M5		E5		M6		E6		M7		E7		M8
jan 2013	M3en M4	M3	ank33		ank34	M4	ank44															
juni 2013	E3 en E4		ank33	E3	ank34		ank44	E4	ank45													
jan 2014	M5								ank45	M5	ank55											
juni 2014	E5										ank55	E5	ank56									
jan 2015	M6												ank56	M6	ank66							
juni 2015	E6														ank66	E6	ank67					
jan 2016	M7																ank67	M7	ank77			
juni 2016	E7																		ank77	E7	ank78	
jan 2017	M8																				ank78	M8

Figuur 4.1 Macrodesign LVS-toetsen Spelling 3.0

De items die voor de overlap of verankering zorgen, duiden we in het macro-design aan met ank, gevolgd door 2 cijfers. Zo duidt ank34 de groep items aan die enerzijds bestaat uit items geconstrueerd voor E3 en anderzijds uit items geconstrueerd voor M4. Die items zijn dus zowel eind groep 3 als medio groep 4 afgenomen. De groep items ank44 bevat items voor M4 en E4, die dus zowel medio groep 4 als eind groep 4 zijn afgenomen. Een item kan hoogstens in één (overlap)groep voorkomen, dat wil zeggen: de ank-blokjes hebben geen gemeenschappelijke items met elkaar en ook niet met de reguliere blokjes M4, E4, M5, E5, M6, E6, M7, E7 en M8.

Longitudinale opzet
Een volledig longitudinaal design impliceert dat een cohort leerlingen gevolgd wordt van M3 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 1 en zal eveneens deelnemen aan de opvolgende momenten M5 (schooljaar 2) en E5 (schooljaar 2). School B zal starten op moment M5 (schooljaar 1) en zal eveneens deelnemen aan de opvolgende momenten E5 (schooljaar 1) en M6 (schooljaar 2). 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 volgende afnamemoment. In het geval van leerjaar 3, met afnamemomenten M3 en E3, vindt de kalibratie plaats op basis van afnamemomenten M3, E3 en M4. Anders dan bij de andere leerjaren vindt hier nog geen ankering naar beneden plaats, aangezien M3 het eerste afnamemoment is. Deze opzet sluit aan bij de inhoudelijke kenmerken van de aangeboden opgaven; een sterke leerling in groep 3 zal wel opgaven uit groep 4 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 macrodesign uit figuur 4.1 worden gefocust. In het geval van groep 3 betreft het dus het gedeelte dat in de figuur hieronder is weergegeven.

Figuur 4.2 Gedeelte macrodesign waarop kalibratie leerjaar 3 is gebaseerd

Jan 2013	M3	ank33		
Juni 2013		ank33	E3	ank34

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.

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 Spelling 3.0. 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 in het onderzoek dat ten grondslag ligt aan de normering sprake is van low stakes afnamesituaties (Keuning et al., 2014). Bij de ontwikkeling van de toetsen Spelling 3.0 is geprobeerd om bias in de normen te vermijden door de afnamesituatie waarin de toets wordt afgenomen tijdens het normeringsonderzoek zoveel mogelijk te laten lijken op de afnamesituatie na uitgave. Er is gekozen voor een *embedded field* onderzoek, waarin nieuw ontwikkelde items voor de derde generatie van het Cito Volgsysteem primair en speciaal onderwijs (3.0) meeliepen in de al bestaande en op scholen toegepaste toetscyclus. Aan afname van de Starttaak van toetsen Spelling uit de tweede generatie LVS-toetsen van Cito zijn eenmalig twee taken met nieuw materiaal toegevoegd.
In totaal lieten scholen hun leerlingen drie taken maken: een taak uit de tweede generatie van de LVS-toetsen en twee taken met nieuw materiaal voor de derde generatie. Deze drie taken zaten samen in één boekje en telden alle drie mee voor het resultaat van de leerling. Voor de leerlingen was onbekend welke taken 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 was dat de normeringssteekproef aangevuld kon worden met resultaten uit dataretour van de tweede generatie LVS-toetsen (zie Keuning et al., 2014).

In figuur 4.3 en 4.4 wordt het *embedded field* design weergegeven. In de designs kunnen we zien dat er vier toetsversies zijn afgenomen voor M3 en zes voor E3. Elke leerling maakte volgens het design een taak van de LVS-toets Spelling uit de tweede generatie. Daarnaast maakte elke leerling twee taken van de beoogde uitgave Spelling 3.0. In Spelling 3.0 worden ook tussentoetsen aangeboden. Deze tussentoetsen hebben een niveau dat ligt tussen twee toetsen voor opvolgende afnamemomenten en zullen verantwoord worden tezamen met het hogere afnamemoment. Zo kan de tussentoets E3M4 gebruikt worden voor leerlingen op het afnamemoment E3 van wie verwacht wordt dat de E3-toets te makkelijk zal zijn, maar eveneens voor de leerlingen op het afnamemoment M4 van wie verwacht wordt dat de M4-toets te moeilijk zal zijn. De toets E3M4 wordt verantwoord in de verantwoording voor de toetsen Spelling 3.0 voor groep 4. De taken M3 deel 1 en M3 deel 2 vormen tezamen de beoogde uitgave Spelling 3.0 M3. De taken M3E3 deel 1 en M3E3 deel 2 vormen tezamen de beoogde uitgave Spelling 3.0 M3E3. De taken E3 deel 1 en E3 deel 2 vormen tezamen de beoogde uitgave Spelling 3.0 E3. De taken E3M4 deel 1 en E3M4 deel 2 vormen tezamen de beoogde uitgave Spelling 3.0 E3M4. Zoals al aangegeven, zal deze toets worden behandeld in de verantwoording voor de toetsen Spelling 3.0 voor groep 4.

Voor zowel M3, E3 als M3E3 en E3M4 zijn ‘reserve-opgaven’ meegenomen in het normeringsonderzoek: elke taak was verlengd met 5 reserve-opgaven[5]. Deze opgaven maakten deel uit van de toets voor de leerlingen 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 M3, E3, M3E3 en/of E3M4.

[5] Uiteindelijk is besloten het voorziene aantal items per taak (25) naar beneden bij te stellen (20). De uiteindelijke toetsen met 40 opgaven per toets voldoen aan de kwaliteitseisen met betrekking tot de betrouwbaarheid (zie hoofdstuk 5) en passen qua lengte beter bij de kenmerken van leerlingen van groep 3.

Zoals te zien in de designs vormden de toetsen M3 respectievelijk E3 uit de tweede generatie van het LVS een stevig anker tussen de toetsboekjes en werd er tussen de beoogde reguliere en tussentoetsen voor Spelling 3.0 geankerd in de normeringsonderzoeken Spelling 3.0 M3 en E3. Omdat de tussentoetsen telkens werden meegenomen in twee opvolgende normeringsonderzoeken is er ook een ankering over de toetsmomenten heen.

De leerlingen maakten zowel oud als nieuw materiaal. Door deze opzet kon de zogenoemde dataretour van de tweede generatie toetsen worden meegenomen in het vaststellen van de normering voor de uitgave Spelling 3.0 groep 3. Ook kunnen we door deze opzet de normering van de nieuw uit te geven toetsen vergelijken met de normering van de toetsen van de tweede generatie en kan de continuïteit tussen de tweede en derde generatie in beeld worden gebracht (zie hoofdstuk 6 over validiteit).

Figuur 4.3 Design Spelling 3.0 M3

Toets	M3 LVS 2e gen.	M3 deel 1	M3 deel 2	M3E3 deel 1	M3E3 deel 2	Leerlingen	Scholen
1	■	■	■			350	14
2	■		■	■		376	16
3	■			■	■	392	18
4	■	■			■	364	17

Figuur 4.4 Design Spelling 3.0 E3

Toets	E3 LVS 2e gen.	M3E3 deel 1	M3E3 deel 2	E3 deel 1	E3 deel 2	E3M4 deel 1	E3M4 deel 2	Leerlingen	Scholen
1	■	■			■			236	14
2	■		■	■				304	11
3	■			■			■	229	10
4	■				■	■		263	11
5	■		■			■		200	9
6	■	■					■	261	12

In het normeringsonderzoek M3 zijn 145 verschillende items voorgelegd aan 1518 leerlingen van groep 3. Hiervan hebben 1482 leerlingen de volledige taken gemaakt. Dit aantal leerlingen is verdeeld over vier boekjes, zoals aangegeven in de voorlaatste kolom van figuur 4.3. Elk boekje bestond uit 85 opgaven verdeeld over drie taken. De 25 opgaven uit de bestaande taak Spelling van de tweede generatie werden door alle 1482 leerlingen gemaakt. De overige opgaven kwamen in twee boekjes voor en werden gemiddeld door 755 leerlingen gemaakt. Op grond van de gegevens uit de kalibratie van de normerings-onderzoeken is de definitieve selectie van items gemaakt voor de uitgave van de toets Spelling 3.0 M3.

In het normeringsonderzoek E3 zijn 205 verschillende items voorgelegd aan 1525 leerlingen van groep 3. Hiervan hebben 1493 leerlingen de volledige taken gemaakt. Dit aantal leerlingen is verdeeld over zes

boekjes, zoals aangegeven in de voorlaatste kolom van figuur 4.4. Elk boekje bestond uit 85 opgaven verdeeld over drie taken. De 25 opgaven uit de bestaande taak Spelling van de tweede generatie werden door alle 1493 leerlingen gemaakt. De overige opgaven kwamen in twee boekjes voor en werden gemiddeld door 500 leerlingen gemaakt. Op grond van de gegevens uit de kalibratie van de normerings-onderzoeken is de definitieve selectie van items gemaakt voor de uitgave van de toets Spelling 3.0 E3.

4.2.2 De stappen in de kalibratie

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

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

$$s = \sum_{i=1}^{k} a_i x_i \,, \tag{4.1}$$

als een 'afdoende statistiek' (*sufficient statistic*) voor de vaardigheid θ. Dit betekent dat alle informatie in de data met betrekking tot de vaardigheid in deze statistiek aanwezig is. Hiervan wordt gebruikgemaakt bij de statistische toetsen in OPLM. Het basisprincipe van de statistische toetsen in OPLM is dat op grond van de afdoende statistiek s de personen in de data kunnen worden gegroepeerd. En binnen deze groepen kan de verwachte proportie goede antwoorden op een item onder het model, $p(+\mid s)$, 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

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

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

Naast deze M-toetsen is er een algemene itemtoets die de volgende vorm heeft

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

Deze zogenaamde S-toets heeft een χ^2 verdeling onder het model. Als globale modeltoets is de R1c-toets (Glas, 1988) geschikt. Ook de distributie van alle afzonderlijke S-toetsen komt hiervoor in aanmerking. Als we deze S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn op het (0,1) interval.

Kortom, als we afzien van de formeel-statistische achtergrond van de gehanteerde toetsen, kan de kalibratieprocedure als volgt worden samengevat:

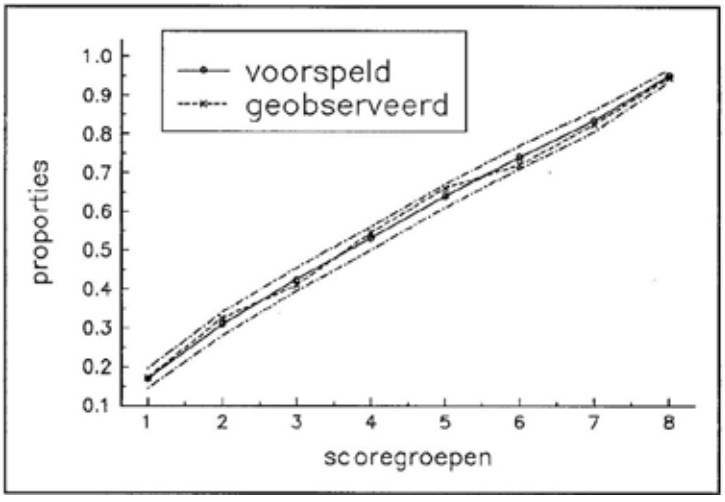
1 Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
2 Vervolgens schatten we de itemparameters met behulp van de CML-methode.
3 Met behulp van de M-toetsen controleren we of de discriminatie-indices goed zijn ingesteld.
4 Een volgende controle betreft de overschrijdingskansen van de S-toetsen en een grafische modelcontrole door middel van het programma 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) score-groepen. 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 parameter-schattingen 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).

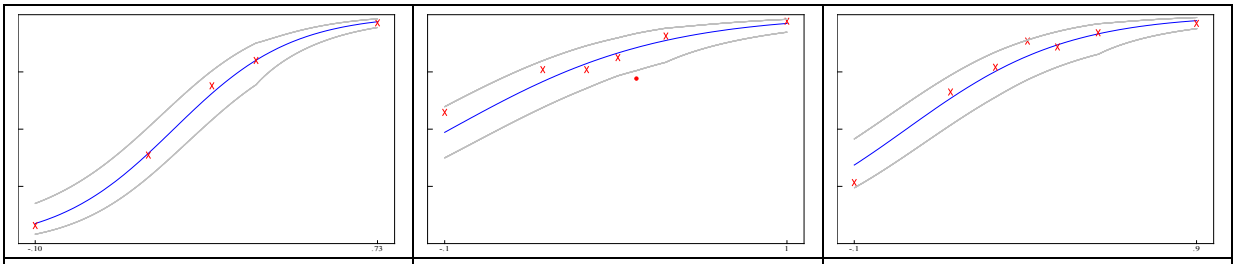
Figuur 4.5 Grafische voorstelling van een Si –toets



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 voor groep 3 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 toetsen Spelling 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 unidimensionaal concept.

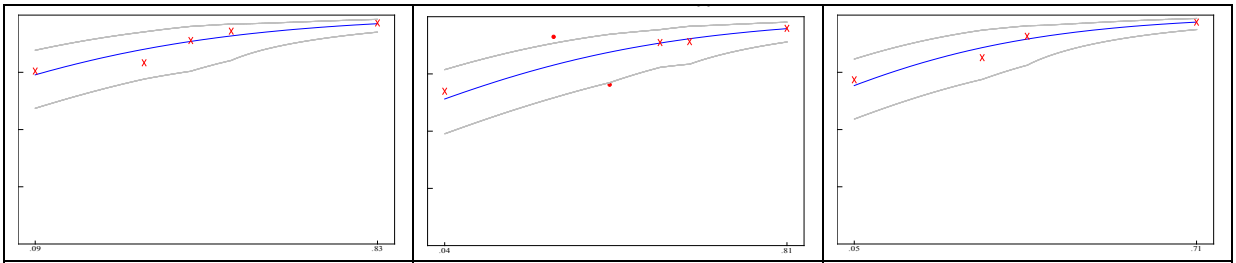
Figuur 4.6 Voorbeelden van S-toetsen voor de toetsen Spelling 3.0 groep 3 met de best passende, de slechtst passende en een qua passing representatieve opgave

M3



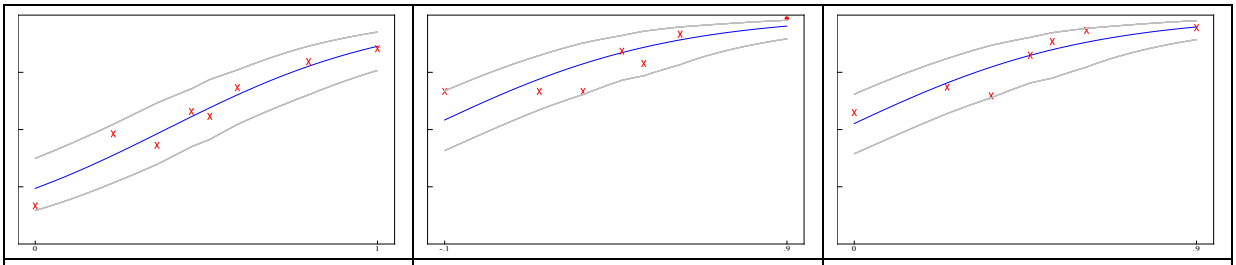
Best passend	Slechtst passend	Representatieve passing

M3E3



Best passend	Slechtst passend	Representatieve passing

E3



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 Spelling 3.0 voor groep 3. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan 0,01, respectievelijk 0,05. Het is duidelijk dat voor de toets 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 (in 109 van de 120 gevallen). Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensionaal construct representeren.

Tabel 4.1 Verdeling van overschrijdingskansen bij S-toetsen voor toetsen Spelling 3.0 groep 3

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.										
M3	1	5	0	6	3	2	4	3	3	4	3	6
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.										
M3E3	1	1	0	4	8	6	4	7	0	4	3	2
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.										
E3	1	2	2	3	7	4	2	4	4	4	5	2

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). Bij steekproeven van deze omvang is alleen laatstgenoemde vuistregel van belang. De modelpassing van de toetsen voldoet aan deze vuistregel. Voor de momenten M3, M3E3 en E3 geldt dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt.

Tabel 4.2 R1c-waarden voor M3, M3E3 en E3

Toetsversie	**R1c**	**df**	**p**
M3	357,4	264	<0,005
M3E3	266,9	244	0,15
E3	309,4	265	0,03

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 vaardigheids-verdeling 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 toetsitems weergegeven. De gemiddelde waarde van de constante is uitstekend te noemen. Voor geen enkele opgave is c groter dan .20.
De conclusie mag luiden dat we ook op basis van deze analyse de kalibratie geslaagd kunnen noemen.

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

Toetsmoment	Constante 'c'	
	Range	Gemiddelde
M3	0,040-0,160	0,076
E3	0,053-0,185	0,106

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen Spelling 3.0 groep 3 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 Spelling 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 Spelling: kan het unidimensionale concept onder de opgaven in de opgavenbank Spelling inderdaad worden opgevat als de vaardigheid spelling? Een geslaagde kalibratie op een unidimensionaal 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 Spelling. De werkwijze die we hieronder beschrijven, komt uit Keuning et al. (2014). Allereerst besteden we aandacht aan de opzet van het normeringsonderzoek, de gehanteerde procedures en de aantallen leerlingen per afnamemoment (paragraaf 4.3.1). Vervolgens komt in paragraaf 4.3.2 de representativiteit van de normsteekproeven aan de orde. De paragraaf wordt afgerond met een presentatie van de resultaten van de normering (i.e. de kenmerken van de vaardigheids-verdelingen op de onderscheiden afnamemomenten; paragraaf 4.3.3).

4.3.1 Opzet

Tijdens het *embedded field* normeringsonderzoek zoals omschreven in paragraaf 4.2.1 werden 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 Spelling van de tweede generatie is bijzonder hoog: de toetsen worden door 88% tot 93% van de scholen gebruikt. Voor de deelnemende scholen aan het normeringsonderzoek is nagegaan of zij als groep afweken van wat men voor de totale populatie van scholen zou mogen verwachten. Niet alleen de gemiddelde score op de Cito Eindtoets Basisonderwijs, maar ook specifiek de score voor spelling bleek voor de scholen in het normerings-

onderzoek niet af te wijken van het populatiegemiddelde voor de Eindtoets. Dit is voor meerdere jaren onderzocht.

Voor het normeringsonderzoek van de toets M3 waren 1200 leerlingen nodig. De scholen die al hadden meegedaan met een proeftoets voor Spelling, werden uitgesloten. Voor het totaal aantal scholen in Nederland (7168 scholen) is een indeling gemaakt naar LOVS-strata (0 t/m 10 procent, 11 t/m 25 procent, 26 t/m 40 procent en 41 procent of meer gewichtsleerlingen) bij schoolgrootte (meer dan 200 leerlingen dan wel minder dan 200 leerlingen). Dit resulteerde in 8 groepen. Vervolgens zijn clustersteekproeven getrokken op dusdanige wijze dat de 8 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%.
Er zijn 1221 scholen aangeschreven. Hiervan waren 68 scholen met in totaal 1660 leerlingen bereid om deel te nemen. Uiteindelijk hebben 65 scholen met 1518 leerlingen daadwerkelijk meegedaan en konden we de gegevens van 1482 leerlingen gebruiken. Leerlingen met veel missings werden namelijk niet meegenomen bij de kalibratie/normering.

De normeringsgroep voor E3 bestond deels uit herhalingsscholen die ook op het afnamemoment M3 en/of M4 hadden meegedaan aan het normeringsonderzoek. (De meeste scholen deden aan zowel M3 als M4 mee.) Ook nu is er weer dezelfde indeling gemaakt naar LOVS-strata. Voor het normeringsonderzoek E3 zijn uiteindelijk in totaal 77 herhalingsscholen van M3/M4 aangeschreven en 854 extra scholen. Omdat na de eerste aanschrijvingsronde maar 49% van de herhalingsscholen uit het normeringsonderzoek Spelling M3/M4 ook bereid bleek deel te nemen aan het normeringsonderzoek Spelling E3 en minder dan 1% van de scholen uit de aanvullende steekproef bereid bleek deel te nemen aan het normeringsonderzoek, zijn in een tweede wervingsronde 1400 extra scholen aangeschreven. Uiteindelijk resulteerde dit in een deelnamebereidheid van zo'n 49% van de herhalingsscholen en 1% van de overig aangeschreven scholen. In totaal meldden zich 69 scholen aan voor het normeringsonderzoek E3 met in totaal 1588 leerlingen. Uiteindelijk hebben 67 scholen met 1525 leerlingen daadwerkelijk meegedaan en konden we de gegevens van 1493 leerlingen gebruiken.

Voor het bepalen van de normering werden de gegevens uit het normeringsonderzoek aangevuld met gegevens uit Cito-dataretour. 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 van wie het onderwijstype onbekend is;
- leerlingen van scholen die het LVS selectief inzetten. In de hogere leerjaren blijken sommige scholen LVS-toetsen 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 hoorde, werden behouden. Daarnaast werden de scholen verwijderd die ook aan de *embedded field* normeringsonderzoeken deelnamen.

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 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 LVS-toetsen van de derde generatie maakt namelijk verhoudingsgewijs af. Met het oog op de constructie en validering van de derde generatie LVS-toetsen is dit onwenselijk.
Door het *embedded field* normeringsonderzoek een groter gewicht te geven, neemt de hoeveelheid data die volledig in de feitelijke toetssituatie verzameld is 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 normeringssteekproef 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 de derde generatie LVS-toetsen is 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 ervoor 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 Eén 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 Eén 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 was niet mogelijk om expliciet rekening te houden met de variabele *etniciteit*, omdat (a) er geen eenduidige referentiegegevens voor de populatie bekend waren, 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 normeringssteekproef 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 werd 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 de Cito-dataretour toegevoegd aan de dataset met normeringsgegevens. Niet elke school heeft daarbij evenveel kans om geselecteerd te worden. Bij de selectie wordt namelijk rekening gehouden met de regio en de urbanisatiegraad van de school en het aantal achterstandsleerlingen. De kans w_{ijk} dat een school met regio i, urbanisatiegraad j en schooltype k geselecteerd wordt, hangt af van het reeds geselecteerde aantal leerlingen N_{S}, het gewenste aantal leerlingen N_{T}, en het beschikbare aantal leerlingen in Cito-dataretour N_{D}:

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

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_{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_{S} dichterbij het gewenste aantal leerlingen N_{T} komt te liggen.
De oplossing is om N_{T} bij de berekening van de gewichten te vermenigvuldigen met een vrij te kiezen constante C en het algoritme te beëindigen in de eerste iteratie waarbij geldt dat $N_{\mathrm{S}} \geq N_{\mathrm{T}}$. 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 normeringssteekproef bepaald is, wordt binnen elke school een steekproef met teruglegging getrokken. Bij het trekken van de steekproef wordt rekening gehouden met w_q. De trekking wordt beëindigd op het moment dat het geselecteerde leerlingaantal gelijk is aan het oorspronkelijke leerlingaantal. De steekproeftrekking wordt per school uitgevoerd, omdat het met het oog op de schoolnormering noodzakelijk is dat de scholen qua

omvang en samenstelling zoveel mogelijk intact blijven. Dit is ook de reden dat in de eerste stap uitsluitend gehele scholen geselecteerd worden en geen individuele leerlingen.

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

Voorbereiding data normeringsonderzoek

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

Toevoegen data uit Cito-dataretour

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

Check leerlingvariabele sekse

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

Het algoritme is toegepast bij de ontwikkeling van de toetsen Spelling 3.0. 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 3 ingesteld op $N_{\mathrm{T}} = 2 \times 1482 = 2964$ en voor afnamemoment einde groep 3 op $N_{\mathrm{T}} = 2 \times 1493 = 2986$.

In tabel 4.4 is te zien in welke aantallen scholen en leerlingen het selectiealgoritme heeft geresulteerd. De conclusie is dat het zowel voor afnamemoment medio groep 3 als voor afnamemoment eind groep 3 tot de gewenste oplossing heeft geleid. De aantallen leerlingen die via het *embedded field* normerings-onderzoek 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 wordt weergegeven welke aantallen leerlingen van de steekproef en van dataretour uiteindelijk zijn meegenomen in de normering.

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

Afnamemoment	Aantal leerlingen			Aantal scholen
	Steekproef	Dataretour	Totaal normering	Normering
M3	1482	1542	2964	133
E3	1493	1530	2996	132

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

4.3.2 Representativiteit

Door de werkwijze die werd gevolgd tijdens de normering is representativiteit van de normerings-steekproeven in principe gegarandeerd. Niettemin werd er een controle uitgevoerd op de representativiteit door de populatieverdelingen verkregen uit gegevens van DUO te vergelijken met de steekproef-verdelingen. In de tabellen 4.5 t/m 4.8 worden de resultaten van de representativiteitsanalyses getoond. De steekproef is geanalyseerd in relatie tot de variabelen stratum (schooltype), regio, urbanisatiegraad en geslacht.

Tabel 4.5 Aantal en percentage leerlingen in de populatie en de steekproef naar stratum

	Populatie	Steekproef			
stratum	%	M3	%	E3	%
0 – 10%	64,8	1921	64,8	1907	64,3
10 – 25%	24,9	738	24,9	758	25,6
25 – 40%	5,2	154	5,2	146	4,9
> 40%	5,1	151	5,1	153	5,2

M3 $\chi^2(3, N = 2964) = 11{,}968$; $p = 0{,}007$; $\phi = 0{,}064$
E3 $\chi^2(3, N = 2964) = 13{,}167$; $p = 0{,}004$; $\phi = 0{,}067$

Tabel 4.6 Aantal en percentage leerlingen in de populatie en de steekproef naar regio

	Populatie	Steekproef			
regio	**%**	**M3**	**%**	**E3**	**%**
Noord	10,1	289	9,8	280	9,5
Oost	22,6	649	21,9	652	22,0
West	47,4	1453	49,0	1449	48,9
Zuid	19,9	573	19,3	583	19,7

M3 $\chi^2(3, N = 2964) = 3{,}270$; $p = 0{,}352$; $\phi = 0{,}033$
E3 $\chi^2(3, N = 2964) = 3{,}365$; $p = 0{,}339$; $\phi = 0{,}034$

Tabel 4.7 Aantal en percentage leerlingen in de populatie en de steekproef naar urbanisatiegraad

	Populatie	Steekproef			
Urbanisatie	**%**	**M3**	**%**	**E3**	**%**
Stad	55,8	1658	55,9	1668	56,3
Land	44,2	1306	44,1	1296	43,7

M3 $\chi^2(1, N = 2964) = 0{,}020$; $p = 0{,}888$; $\phi = 0{,}003$
E3 $\chi^2(1, N = 2964) = 0{,}263$; $p = 0{,}608$; $\phi = 0{,}009$

Tabel 4.8 Aantal en percentage leerlingen in de populatie en de steekproef naar geslacht

	Populatie	Steekproef			
Geslacht	**%**	**M3**	**%**	**E3**	**%**
jongen	50,4	1429	51,6	1410	52,3
meisje	49,6	1343	48,4	1284	47,7

M3 $\chi^2(1, N = 2772) = 1{,}415$; $p = 0{,}234$; $\phi = 0{,}023$
E3 $\chi^2(1, N = 2694) = 3{,}960$; $p = 0{,}047$; $\phi = 0{,}038$

De χ^2-waarden zijn laag en in slechts een paar gevallen significant. Bij grotere steekproeven zegt significantie echter niet zoveel. Het is beter om de effectgrootte ϕ als uitgangspunt te nemen.

Formule 4.1 Berekening van de effectgrootte ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

We zien dat de effectgroottes ver onder de 0,10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). Ze zijn met 0,067 nog het hoogst voor de variabele *stratum* bij het afnamemoment E3. De conclusie is niettemin dat de normeringssteekproeven een zeer goede afspiegeling vormen van de populatie.

4.3.3 Normeringsresultaten

Na de hierboven beschreven procedure doorlopen te hebben en de normeringssteekproef te hebben samengesteld, kon de normering worden bepaald. Naast het gemiddelde werden de percentielen bepaald. Dat gebeurde op basis van de verdeling van scores die werden gevonden in de normeringssteekproef zoals die is samengesteld op basis van het *embedded field* normeringsonderzoek en Cito-dataretour. Om de scores van leerlingen te kunnen vergelijken over de tijd worden vaardigheidsscores gebruikt. Uit de ruwe scores van de leerlingen uit het *embedded field* normeringsonderzoek en Cito-dataretour werden “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., 2014). De normering werd vervolgens gebaseerd op de “plausible values” van de leerlingen in de normeringssteekproef. In paragraaf 3.3 is de verdeling van “plausible values” voor afnamemomenten M3-E3 te zien. De ”plausible values” voor deze afnamemomenten vormen een normale verdeling. Op basis van deze scoreverdeling werden 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. Tabel 4.6 geeft de norm-gegevens voor de toetsen Spelling 3.0 groep 3.

Tabel 4.6 Normtabel op leerlingniveau voor Spelling 3.0 groep 3

Tijd	***M***	***SD***	***K***	***S***	***P*10**	***P*20**	***P*25**	***P*40**	***P*50**	***P*60**	***P*75**	***P*80**
M3	150.0	50,0	1,045	-0,380	89,0	113,8	122,0	141,4	152,1	162,6	181,4	188,9
E3	200.4	40,4	0,631	-0,142	151,6	167,6	173,8	190,2	200,2	210,4	227,0	233,8

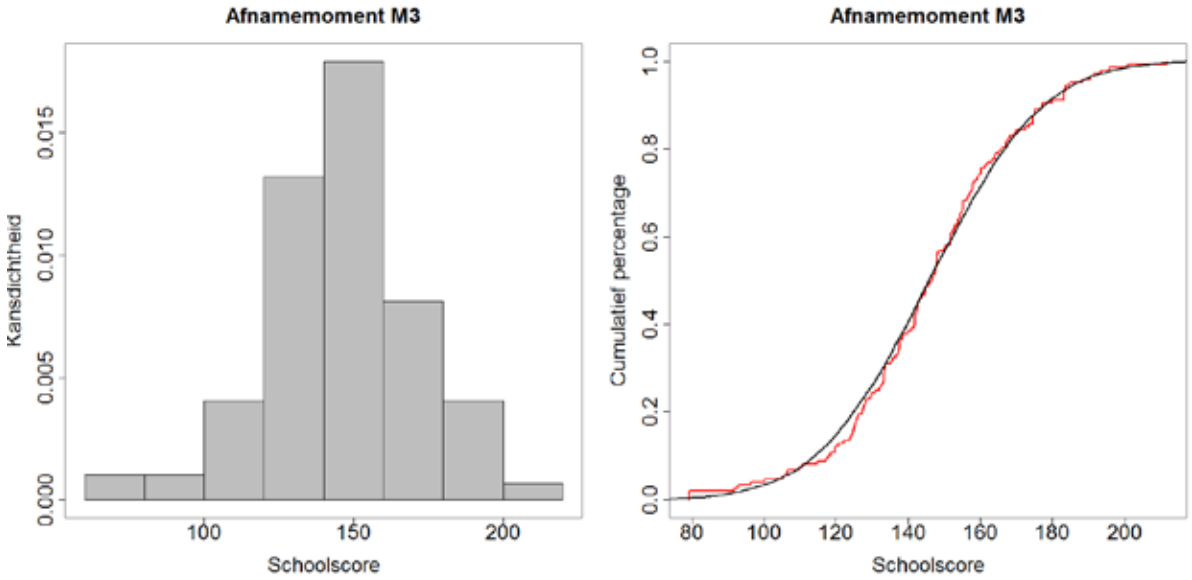
Naast een normering op leerlingniveau kent Cito ook een normering op schoolniveau. Om de school-verdeling te bepalen werd het intercept-only multilevelmodel gebruikt met een gemiddelde per school en een variantie op school- en leerlingniveau. De schatting van het model verloopt via een bootstrap procedure. Dit betekent dat het multilevelmodel 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 multilevelmodel geschat en de intraklassecorrelatie en het designeffect uitgerekend. Tabel 4.7 laat de resultaten van de bootstrapprocedure zien. We zien dat de uitkomsten behoorlijk stabiel zijn. In onderwijskundig onderzoek liggen de intraklassecorrelaties doorgaans tussen de 0,05 en 0,25 (Snijders & Bosker, 1993; Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Schochet, 2008). Als vuistregel wordt vaak aangehouden dat een multilevel analyse zinvol is als de intraklassecorrelatie 0,04 of meer bedraagt. Dit is hier het geval.

Tabel 4.7 Samenvatting uitkomsten multilevel analyse Spelling 3.0 groep 3

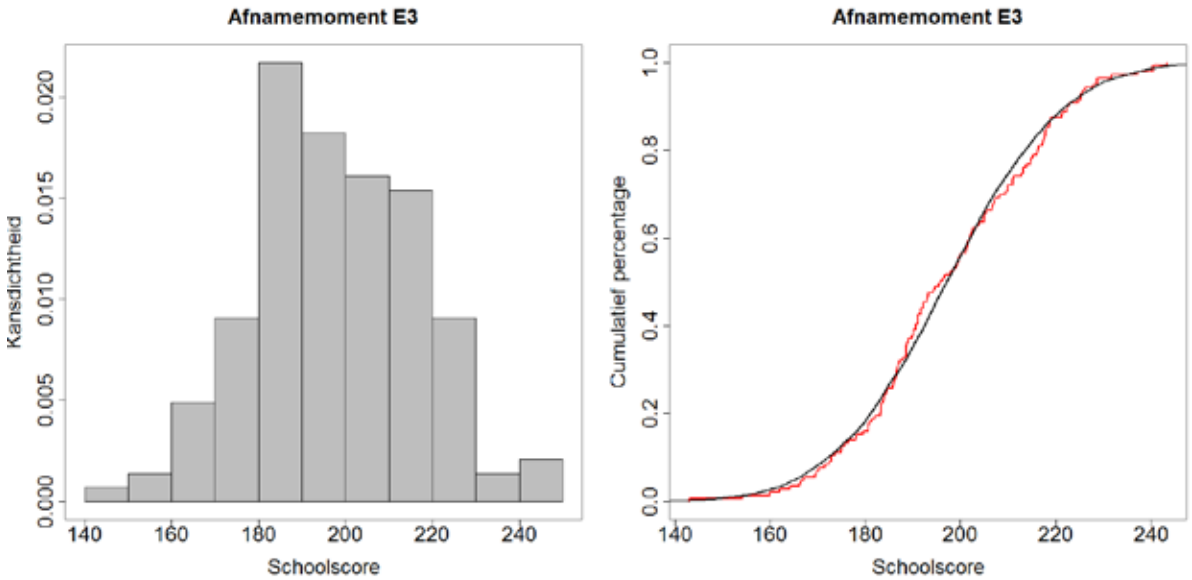
Moment	**Aantal replicaties**	**Aantal scholen**	**Gemiddelde**	**SD School**	**SD Leerling**	**ICC**
M3	20	148	145,1	21,6	45,4	0,19
E3	20	143	197,8	19,1	38,3	0,20

Figuren 4.7 en 4.8 laten de verdelingen van schoolgemiddelden zien. Het is lastig te bepalen of de schoolgemiddelden een normale verdeling volgen met een scholenaantal van 161. Op het eerste gezicht lijkt er sprake te zijn van een normale verdeling. Op basis van de resultaten van de bootstrapprocedure zijn de percentielen voor de vaardigheidsverdeling A tot en met E en I tot en met V berekend. Tabel 4.8 geeft de normgegevens op schoolniveau. De percentielen komen dichter bij elkaar te liggen dan in de leerling-verdeling. De afstanden zijn echter nog wel groot genoeg om scholen zinvol te classificeren in de verschillende niveaus.

Figuur 4.7 Verdeling van de schoolgemiddelden voor Spelling 3.0 M3



Figuur 4.8 Verdeling van de schoolgemiddelden voor Spelling 3.0 E3



Tabel 4.8 Normtabel op schoolniveau voor Spelling 3.0 groep 3

Tijd	*M*	*SD*	*P*10	*P*20	*P*25	*P*40	*P*50	*P*60	*P*75	*P*80
M3	145,1	21,6	117,4	126,9	130,5	139,7	145,1	150,6	159,7	163,3
E3	197,8	19,1	173,3	181,7	184,8	192,9	197,8	202,6	210,7	213,9

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 Spelling 3.0 groep 3 een geldigheid aanhouden tot en met 2023. Daarnaast monitort Cito periodiek de normering: jaarlijks wordt aan de hand van representatieve afnamedata nagegaan of er systematisch verschuivingen in het prestatieniveau plaatsvinden. Indien nodig wordt de normering aangepast.

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Betrouwbaarheid

In hoofdstuk 4 is onder meer aangegeven dat de meeste leerlingen die deelgenomen hebben aan het normeringsonderzoek slechts een deel van de items gemaakt hebben die uiteindelijk in de toetsen Spelling 3.0 opgenomen zijn. De betrouwbaarheid van de toetsen in klassieke zin is dan ook niet rechtstreeks te bepalen. Het is echter wel mogelijk om de betrouwbaarheid van iedere toets te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de toetsen OPLM-geschaald zijn. Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het feit dat de toetsen volledig bestaan uit OPLM-gekalibreerde items. Om relevante beschrijvende gegevens bij de verschillende toetsen te genereren, is gebruikgemaakt van het programma OPLAT (Verhelst, Glas en Verstralen, 1995).

In OPLAT wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ. Deze verwachte waarde wordt aangeduid met 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 Spelling. In de tweede kolom staat de maximumscore, voor iedere toets is deze gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. De derde kolom geeft de geschatte gemiddelde scores van de leerlingen op de verschillende toetsen. De vierde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van iedere toets. De vijfde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de verschillende toetsen (of toetsonderdelen) is.

Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen Spelling 3.0 uit het Cito Volgsysteem) geeft de COTAN (COmmissie TestAangelegenheden Nederland van het Nederlands Instituut van Psychologen) aan dat een betrouwbaarheidscoëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 voldoende, en een

betrouwbaarheidscoëfficiënt hoger dan 0,80 goed (Evers, Lucassen, Meijer en Sijtsma 2010, p. 33). Op grond van dit criterium is de meetnauwkeurigheid van de toetsen goed te noemen.

Tabel 5.1 Beschrijvende gegevens bij de toetsen Spelling groep 3

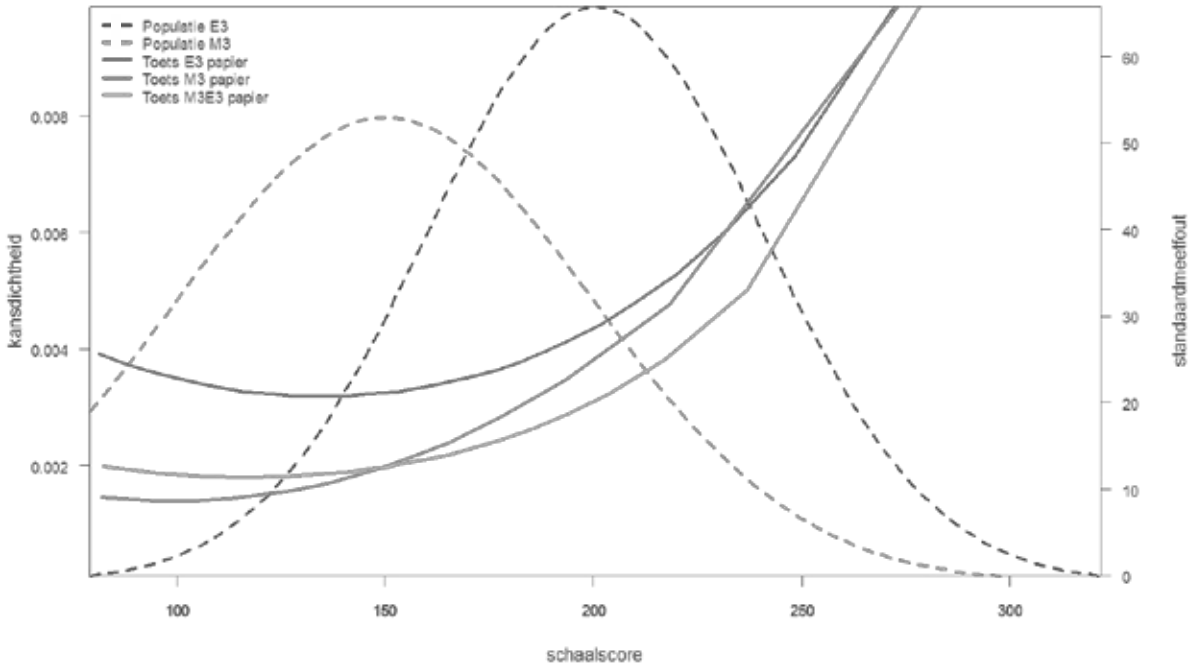
Toets	Maximum score	Gemiddelde	Standaard-meetfout	MAcc	Test-hertest (simulatie)
M3	40	31,2	2,226	0,92	0,92
M3E3	40	34,8	1,909	0,86	0,86
E3	40	32,0	2,281	0,86	0,86

Er heeft geen test-hertestonderzoek plaatsgevonden. De afnamecontexten van de LVS-toetsen Spelling 3.0 lenen 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 sprake van invloed van een test-hertestinterval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden. De resultaten zijn weergegeven in tabel 5.1 (zie kolom 6). De uitkomst komt exact overeen met de eerder berekende coëfficiënten en leiden dan ook tot dezelfde conclusie met betrekking tot de betrouwbaarheid van de toetsen Spelling 3.0.

5.2 Nauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen en geven geen beeld van de lokale meetnauwkeurigheid van de toetsen Spelling 3.0 voor groep 3. Figuur 5.1 geeft grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid van deze toetsen. In deze figuur staat voor deze toetsen de grootte van de meetfout op de vaardigheidsschaal afgebeeld. Ook is de kansdichtheidfunctie voor de normgroep op het afname-moment opgenomen. Deze laat zien hoe de vaardigheid van de leerlingen verdeeld is over de vaardigheidsschaal in de populatie die de toets gemaakt heeft. De figuur maakt 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 toetsen M3-E3 en de kansdichtheidsfuncties voor de M3 en E3-populatie



Betrouwbaarheidstabellen

De betekenis van de (lokale) meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit betrouwbaarheidstabellen. Tabellen 5.2 tot en met 5.4 laten voor de toetsen M3, M3E3 en E3, respectievelijk afnamemomenten medio groep 3 en einde groep 3, zien hoe vaak de werkelijke vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat Tabel 5.2 zien dat 89,6 procent van de leerlingen die halverwege groep 3 op basis van de M3-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 90 procent. Verder laat de linkerkant van Tabel 5.5a zien dat 10,4 procent van de leerlingen in scoregroep V een vaardigheidsscore heeft die in werkelijkheid in scoregroep IV valt. De overige getallen in Tabellen 5.2 tot en met 5.4 zijn op dezelfde wijze te interpreteren.

Tabel 5.2 Betrouwbaarheidstabel toets M3 voor afnamemoment medio 3

	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	**89,6**	10,4	0,1	0,0	0,0	E	**89,8**	10,2	0,0	0,0	0,0
IV	15,8	**65,1**	17,9	1,3	0,0	D	10,2	**72,7**	17,0	0,1	0,0
III	1,0	26,4	**46,0**	24,6	2,0	C	0,0	16,4	**64,3**	18,7	0,1
II	0,2	6,5	23,7	**43,4**	26,2	B	0,0	1,4	25,1	**51,2**	22,4
I	0,3	2,0	6,3	19,0	**72,3**	A	0,0	0,1	5,1	20,0	**74,4**

Tabel 5.3 Betrouwbaarheidstabel Toets M3E3 voor afnamemoment einde 3

	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	**83,6**	15,5	0,9	0,0	0,0	E	**83,2**	16,1	0,7	0,0	0,0
IV	24,4	**50,1**	21,7	3,7	0,1	D	20,8	**53,6**	24,7	0,8	0,0
III	4,9	26,5	**37,7**	25,2	5,6	C	2,2	21,0	**53,1**	21,7	2,0
II	1,6	9,8	22,8	**34,3**	31,4	B	0,3	4,0	25,7	**42,6**	27,4
I	1,4	4,0	8,7	17,9	**68,0**	A	0,5	1,6	7,9	20,2	**69,9**

Tabel 5.4 Betrouwbaarheidstabel Toets E3 voor afnamemoment einde 3

	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	**83,4**	15,8	0,8	0,0	0,0	E	**82,4**	16,7	0,8	0,0	0,0
IV	23,0	**52,5**	21,6	2,8	0,1	D	20,4	**54,3**	24,7	0,7	0,0
III	3,1	26,1	**41,9**	25,1	3,7	C	1,6	19,8	**56,4**	21,0	1,2
II	0,6	7,1	23,8	**40,0**	28,6	B	0,1	2,3	24,7	**48,2**	24,7
I	0,2	1,7	6,6	19,1	**72,4**	A	0,0	0,4	4 ,8	20,4	**74,4**

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 3 en einde groep 3 zijn te vinden in Tabel 5.5. 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 5.5 aannemelijk dat de uitkomsten duidelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969). Gemiddeld gezien scoort, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 92 tot 98 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 55 tot 67 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore

bij beide afnamemomenten gemiddeld gezien in ruim 50 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. De resultaten stemmen hiermee tot tevredenheid: het percentage misclassificaties is beperkt. De laagste waarden zien we bij toets M3E3, en dan met name bij de hoogste scoregroep. Dat is conform verwachting, aangezien deze toets – die wat makkelijker is dan de toets E3 – expliciet bedoeld is voor de minst vaardige leerlingen aan het eind van groep 3. 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 Spelling 3.0 groep 3 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 1 niveau verschil.

Tabel 5.5 Samenvattende indices toetsen M3, M3E3 en E3 op afnamemomenten groep 3

	Toets M3, afnamemoment M3		**Toets M3E3, afnamemoment E3**		**Toets E3, afnamemoment E3**	
	score-groep I t/m V	**score-groep A t/m E**	**score-groep I t/m V**	**score-groep A t/m E**	**score-groep I t/m V**	**score-groep A t/m E**
Marginal classification accuracy	63,3	67,4	54,9	57,7	58,2	61,1
Accuracy plus/minus 1 niveau	96,0	98,1	91,9	95,2	94,7	97,2

6 Validiteit

De begripsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de toetsscores toe te schrijven zijn aan verklarende concepten en constructen die deel uitmaken van het theoretische kader dat aan de ontwikkeling van de toets ten grondslag ligt. Bij leervorderingstoetsen, zoals deze toets Spelling 3.0 voor groep 3, 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 uiteraard 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 unidimensionale 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 over de convergente en divergente validiteit. Als laatste komen in paragraaf 6.6 verschillen tussen relevante subgroepen aan bod.

6.1 Inhoudsvaliditeit

De samenstelling van de toetsen is bepaald door inhoudelijke criteria en psychometrische criteria. Voor de inhoudsvaliditeit zijn de inhoudelijke criteria relevant. Inhoudelijk zijn richtinggevend geweest het referentiekader Nederlandse taal (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a), de kerndoelen Nederlandse taal (Ministerie van OCW, 2006), de Leerstoflijnen Begrippenlijst en taalverzorging (Van der Beek & Paus, 2011) en recente wetenschappelijke publicaties over spelling. Deze bronnen, tezamen met een uitgebreide methodenanalyse, vormden de basis voor de domeinbeschrijving van de toetsen Spelling. In hoofdstuk 3 is de domeinbeschrijving uitgewerkt in een beschrijving en verantwoording van spellingcategorieën binnen de toetsen Spelling. De constructie van de opgaven is afgeleid van deze domeinindeling en ook de definitieve selectie van opgaven in de toetsen is gebaseerd op een gewenste verdeling van opgaven binnen en over de verschillende spellingcategorieën. Beoogd is de toetsen onafhankelijk samen te stellen van de verschillende onderwijsmethoden in die zin dat de getoetste stof in het merendeel van de methodes aan bod is gekomen en de inhoud van de toetsen niet in meerdere mate tot uitdrukking komt in een van de methoden. Bij de constructie van de opgaven zijn leerkrachten uit het onderwijs betrokken zodat de opgaven voor wat betreft spellingmoeilijkheid en voor wat betreft context aansluiten bij het ontwikkelingsniveau van leerlingen van groep 3.
Kortom, de gewenste toetsinhoud – in termen van beoogde aantallen items voor de onderscheiden spellingcategorieën – is beargumenteerd op basis van wetenschappelijk goed verdedigbare keuzes omtrent referentiekader, kerndoelen en leerstoflijnen en in overeenstemming gebracht met de meest gangbare lesmethoden. De opgaven zijn geconstrueerd door leerkrachten uit het basisonderwijs, van een correcte context voorzien en empirisch uitgetest in proef- en normeringsonderzoeken. De gewenste verdeling over spellingcategorieën, ten slotte, is in de uiteindelijke itemselectie daadwerkelijk gerealiseerd. Dit alles vormt een degelijke basis voor de inhoudsvaliditeit van de toetsen.

6.2 Unidimensionaliteit, respectievelijk structuur

Zoals in hoofdstuk 4 al aangegeven is, 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 van deze statistische toetsingen 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 unidimensionaal 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 belangrijkste vuistregel dat R1c bij voorkeur niet groter zou moeten zijn dan ongeveer anderhalf maal het aantal vrijheidsgraden. In tabel 4.2 zijn deze waarden te vinden. De modelpassing van de toetsen M3, E3 en M3E3 voldoet aan deze vuistregel.

Unidimensionaliteit van de te meten vaardigheid is het uitgangspunt van het meetmodel van de toetsen Spelling 3.0. Ondanks de inhoudelijk relevante indeling van de opgaven is een unidimensionale schaal gerealiseerd. Dit betekent dat met elke willekeurige subset van items dezelfde onderliggende vaardigheid kan worden vastgesteld.

Ten slotte kan de nauwkeurigheid van de itemparameterschattingen aan de hand van de constante ‘c’ (zie hierover het COTAN Beoordelingssysteem; Evers, Lucassen, Meijer & Sijtsma, 2010, p 40) als uitstekend worden beoordeeld. In hoofdstuk 4 is deze informatie al weergegeven maar omdat deze ook relevant is voor de validiteit wordt deze hier nog eens aangehaald. In tabel 4.3 zijn gemiddelde en range van deze waarden voor alle toetsitems weergegeven. De gemiddelde waarden van de constante zijn te interpreteren als uitstekend. Voor geen enkele opgave is c bovendien groter dan 0,20. De conclusie mag luiden dat we ook op basis van deze analyse de kalibratie geslaagd kunnen noemen.

6.3 Itemkwaliteit

In tabel 6.2 zijn de ranges en de gemiddelden weergegeven voor de p-waarden en de Rit-waarden van de items van de toetsen M3, M3E3 en E3. Bij alle toetsen is te zien dat de p-waarden liggen tussen de 0,44 en 0,95. Er is gestreefd naar p-waarden van de items tussen 0,40 en de 0,90. Enkele uitzonderingen daargelaten is dat gelukt. Er is gestreefd naar een goede spreiding van moeilijkheid over de items.
De gemiddelde moeilijkheid van de M3-, M3E3- en E3-toets ligt tussen de 0,78 en 0,87. Dat is iets hoger dan de gemiddelde p-waarde tussen de 0,65 en 0,75 waarnaar in het algemeen bij LVS-toetsen wordt gestreefd. Daarmee zijn de toetsen niet te moeilijk en wordt voorkomen dat de leerling gefrustreerd raakt tijdens de toetsafname.

Tabel 6.2 Range en gemiddelde van p- en Rit-waarden voor de toetsen M3, M3E3 en E3

	P-waarden		**Rit-waarden**		**N items**
	Range	**Gemiddelde**	**Range**	**Gemiddeld**	
M3	0,44 - 0,92	0,78	0,24 - 0,67	0,44	40
M3E3	0,66 - 0,95	0,87	0,20 - 0,47	0,35	40
E3	0,46 - 0,89	0,80	0,24 - 0,49	0,36	40

Het feit dat de toetsen Spelling 3.0 voor groep 3 relatief gemakkelijk zijn, hangt ook samen met het feit dat een aantal leerlingen in groep 3 al veel meer kan spellen dan ze volgens de gehanteerde methode geleerd hebben op school. De toets moeilijker maken kan alleen door veel dicteewoorden toe te voegen van categorieën die de leerling nog niet heeft gehad in het onderwijs. Omdat we de toetsen wilden laten aansluiten bij het onderwijsaanbod – en dat ook de wens is van de scholen – hebben we dat niet gedaan. Voor de toets M3 geldt wel dat die is verrijkt met dicteewoorden uit categorie 2, terwijl die categorie medio groep 3 nog niet is aangeboden. Zelfs na het toevoegen van deze dicteewoorden en het inkorten van de toetsen door de allermakkelijkste items na de proeftoetsing niet op te nemen zijn de toetsen voor groep 3 voor de meeste leerlingen nog heel gemakkelijk. Ze geven daarom beperktere informatie over de meer vaardige leerlingen: de leerlingen die gemiddeld tot zeer sterk zijn in spelling kunnen zich met deze toetsen

minder (goed) onderscheiden van de gemiddelde leerling. De leerlingen die zwak zijn in spelling zijn wel goed te signaleren door afname van de toetsen voor groep 3.
Bij geen enkele toets ligt de Rit-waarde onder de 0,20. De gemiddelde Rit-waarde is voor alle drie de toetsen 0,35 of hoger. Door de Cotan wordt een Rit-waarde hoger dan 0,30 gekwalificeerd als goed. Met een gemiddelde van 0,35 of hoger is de itemkwaliteit van de toetsen goed te noemen. Bijlage 1 bevat een volledig overzicht van de p-waarden en de Rit-waarden van de items van de toetsen.

In tabel 6.3 zijn de verdelingskarakteristieken gegeven van de ruwe scores op de verschillende toetsen. De gemiddelden komen uiteraard overeen met wat men bij een gegeven aantal items mag verwachten bij de gekozen (gemiddelde) moeilijkheidsgraad. Omdat deze gemiddelde moeilijkheidsgraad voor alle onderdelen rond 0,80 ligt, zijn de verdelingen linksscheef (vergelijk de negatieve waarden in de kolom 'skewness'), de ene wat meer dan de andere. De verdelingen zijn ééntoppig. De toppen van de verdelingen van de toetsen lijken vrij sterk op elkaar, al is die van M3E3 wat gepiekter.

Tabel 6.3 Verdelingskenmerken van de toetsen Spelling groep 3

Toets	Aantal opgaven	Gemiddelde	SD	Skewness	Kurtosis
M3	40	31,2	7,65	-1,14	0,69
M3E3	40	34,8	5,08	-1,67	3,35
E3	40	32,0	6,12	-1,16	1,23

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). Bij de toetsen M3 en E3 was deze S-statistiek bij geen enkele toetsopgave significant (bij $\alpha = 0,01$). Er is in deze toetsen dus geen sprake van DIF naar sekse. Bij toets M3E3 was dit in geringe mate het geval, namelijk bij vijf van de veertig items. Er was bij deze vijf items geen sprake van een systematische vorm van DIF, in die zin dat niet steeds hoog- of laagscorende jongens dan wel meisjes werden bevoordeeld.

6.5 Soortgenootonderzoek

In het kader van het soortgenootonderzoek is gekeken naar de convergente validiteit door de samenhang te onderzoeken met de voorgaande versie van de tweede generatie van de LVS-toetsen Spelling.
De divergente validiteit is onderzocht door de samenhangen met andere taalonderdelen en rekenen-wiskunde te onderzoeken.

Convergente validiteit: Correlatie scores Spelling 3.0 met LVS-toetsen Spelling 2e generatie
De leerlingen van de normeringssteekproef hebben zowel opgaven van de toetsen Spelling van de tweede generatie als opgaven van Spelling 3.0 gemaakt en daardoor kan een correlatie worden bepaald tussen de score op de toetsen van de tweede en de derde generatie. De latente correlaties waren hoog voor M3 ($r = 0,94$; $N = 1483$;) en E3 ($r = 0,95$; $N = 1494$;). Aangezien de begripsvaliditeit van de toetsen Spelling van de tweede generatie door de Cotan positief beoordeeld is, vormt deze hoge correlatie een belangrijk element in de bewijsvoering voor de validiteit van de toetsen Spelling 3.0.

Divergente validiteit: Correlatie scores LVS 3.0 met diverse toetsen leervorderingen
Aan de scholen die hebben deelgenomen aan het normeringsonderzoek is gevraagd of ze toestemming gaven voor het inzetten van dataretour van de betreffende leerlingen van andere LVS-toetsen uit het Cito

Volgsysteem. Met de dataretour-functie zijn van de leerlingen van het normeringsonderzoek ook de scores op andere leervorderingentoetsen van Cito beschikbaar. Op basis van algemene cognitieve verschillen (intelligentie) is er altijd sprake van een zekere samenhang tussen verschillende vakgebieden. Hoe sterk deze samenhang is, hangt af van het vakgebied. We verwachtten dat de spellingvaardigheid in groep 3 redelijk sterk samenhangt met de vaardigheid in technisch lezen. Deze twee vaardigheden gaan in groep 3 namelijk gelijk op: in groep 3 spellen de leerlingen klankzuivere woorden, die ze ook hebben leren lezen. Leerlingen leren eerst bepaalde woorden lezen en leren ze dan ook spellen. Ook verwachtten we een redelijk sterke samenhang met begrijpend lezen. Leerlingen die goed zijn in begrijpend lezen, zullen in het algemeen meer lezen dan leerlingen die minder goed zijn in begrijpend lezen. Goede lezers in groep 3 lezen ook sneller en krijgen méér woorden onder ogen dan zwakkere lezers. Hoe vaker een leerling een woord gelezen heeft, hoe beter hij in staat zal zijn het woord te spellen. De correlatie met rekenen-wiskunde zal naar verwachting matig zijn, omdat dat een geheel andere vaardigheid betreft dan spelling.
In tabel 6.4 is de correlatie tussen de toets Spelling E3 en de toetsen Rekenen-Wiskunde E3, Begrijpend lezen E3 en Technisch lezen (DMT E3 en Leestempo E3) weergegeven.

Tabel 6.4 Correlaties tussen Spelling E3 en verschillende andere LVS-toetsen

	Spelling E3*	**Aantal leerlingen**
Cito Rekenen-Wiskunde E3	0,47	394
Cito Begrijpend lezen E3	0,61	212
Cito Technisch lezen – DMT E3	0,60	393
Cito Technisch lezen – Leestempo E3	0,68	148

* Deze correlaties zijn gecorrigeerd voor attenuatie.

Uit tabel 6.4 blijkt dat de correlatie tussen enerzijds Spelling en anderzijds Technisch lezen en Begrijpend lezen inderdaad redelijk hoog is. Voor Spelling is de correlatie met Rekenen-Wiskunde lager dan de correlatie met de leestoetsen. Dat is volgens verwachting: de spellingvaardigheid hangt in het begin van de basisschoolperiode samen met de vaardigheid in (technisch en begrijpend) lezen en niet zozeer met Rekenen-Wiskunde.

Samenvattend kan dus gesteld worden dat de correlaties van de toets Spelling 3.0 E3 conform de verwachtingen zijn. De correlatie met de toets Spelling van de tweede generatie, waarvan de begrips-validiteit positief is beoordeeld, is zeer hoog. Deze correlatie is duidelijk hoger dan de correlaties met andere leervorderingentoetsen en vormt daarmee een ondersteuning voor de validiteit van de toets.
De data geven aan dat er gemeten wordt wat men beoogt te meten, namelijk spellingvaardigheid.

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 drie groepen. Jongere leerlingen zijn leerlingen die op 1 oktober in groep 3 nog geen 6 jaar waren, reguliere leerlingen zijn leerlingen die op dat moment tussen de 6 en de 7 jaar waren en de groep oudere leerlingen bestaat uit leerlingen die op 1 oktober in groep 3 ouder dan 7 jaar waren. In tabel 6.5 wordt de gemiddelde score van de verschillende leeftijdsgroepen weergegeven.

Tabel 6.5 Gemiddelde score per leeftijdsgroep voor de afnamemomenten M3 en E3

M3

Groep	Aantal	M	SD	Effectgrootte (*d*)
Jonger	106	162,8	47,4	0,26
Regulier	1140	151,6	45,5	0,01
Ouder	81	128,7	61,9	-0.41

E3

Groep	Aantal	M	SD	Effectgrootte (*d*)
Jonger	126	208,0	39,5	0,20
Regulier	1097	200,1	40,8	0,01
Ouder	80	183,2	50,3	-0,36

Het patroon in tabel 6.5 is naar verwachting en komt overeen met eerder gevonden verschillen tussen reguliere en vertraagde en versnelde leerlingen (zie bijv. Van Til, Van Weerden, Hemker & Keune, 2014). De groepen jongere leerlingen van M3 en E3 scoren hoger dan gemiddeld. Dit is in beide gevallen een klein effect. Deze leerlingen zijn de versnelde leerlingen die op grond van hun cognitieve capaciteiten en/of leerprestaties een groep hebben overgeslagen. Aan de andere kant scoren de oudste groepen leerlingen in elk van de afnamemomenten het laagst. Hier is sprake van een groter, maar nog steeds klein effect. Ook hier is dat naar verwachting, aangezien deze leerlingen in veel gevallen op grond van hun leerprestaties een jaar gedoubleerd hebben. De leerlingen die in de reguliere jaargroep zitten behalen een score die valt tussen de scores van de jongere en oudere leerlingen. Voor alle leerlingen in de eigen jaargroep zijn de scores vergelijkbaar.

Bij de variabele geslacht is de gemiddelde score van jongens en meisjes bepaald, zie tabel 6.6.

Tabel 6.6 Gemiddelde score jongen-meisje

Moment	Geslacht	Aantal	M	SD
M3	jongen	661	146,5	46,7
	meisje	629	152,9	46,0

Moment	Geslacht	Aantal	M	SD
E3	jongen	649	196,9	43,2
	meisje	617	202,2	40,2

Per afnamemoment is in tabel 6.6 de gemiddelde score van jongens en meisjes weergegeven. Op M3 en E3 scoren meisjes hoger dan jongens. Het is een bekend verschijnsel dat meisjes beter scoren op spellingtoetsen dan jongens (zie bijv. Van Til, Van Weerden, Hemker & Keune, 2014). In termen van de effectgrootte Cohens *d* is er sprake van een verwaarloosbaar verschil (m3: $d = 0,14$ en e3: $d = 0,13$).

Ten slotte is er gekeken naar het leerlinggewicht. Per afnamemoment is in tabel 6.7 de gemiddelde score weergegeven van leerlingen naar het zogeheten leerlinggewicht. Voor toekenning van een gewicht is het opleidingsniveau van beide ouders het hoofdcriterium en het wordt verfijnd door twee niveaus te onderscheiden. Hieraan zijn twee gewichten gekoppeld, namelijk 0,3 voor kinderen van ouders met

maximaal onderwijs op lbo/vbo-niveau en 1,2 voor kinderen van ouders met maximaal basisonderwijs. We verwachten dat leerlingen met een leerlinggewicht lager scoren dan leerlingen zonder leerlinggewicht.

Tabel 6.7 Gemiddelde score naar leerlinggewicht

Moment	Gewicht	Aantal	M	SD	Effectgrootte (*d*)
M3	0,0	785	151,9	45,3	0,04
	0,3	52	128,0	49,2	-0,43
	1,2	45	115,3	66,3	-0,61

Moment	Gewicht	Aantal	M	SD	
E3	0,0	943	201,5	40,9	0,03
	0,3	50	187,7	52,8	-0,26
	1,2	55	171,1	51,3	-0,62

In tabel 6.7 is te zien dat leerlingen met een leerlinggewicht van 0,3 en 1,2 op M3 en E3 inderdaad lager scoren dan leerlingen zonder leerlinggewicht (gewicht van nul). In termen van de effectgrootte Cohens *d* is er sprake van een medium effect op afnamemoment M3 voor leerlingen met een gewicht van 0,3 en een klein effect voor deze groep op afnamemoment E3. Voor zowel M3 als E3 is het effect voor leerlingen met een gewicht van 1,2 wat groter, maar ook hier wordt gesproken van een medium effect. Ook deze bevindingen komen overeen met eerder gevonden verschillen: de leerlingen zonder leerlinggewicht scoren het best, de leerlingen met een laag gewicht doen het duidelijk minder en de leerlingen met een hoog gewicht scoren het laagst (Van Til, Van Weerden, Hemker & Keune, 2014).

Waar wij op theoretische gronden verwachtingen hadden over verschillen tussen subgroepen, zijn deze bevestigd. Dit levert een bijdrage aan de validiteit van de toetsen Spelling 3.0.

7 Samenvatting

In dit samenvattende hoofdstuk geven we kort weer wat in de voorafgaande hoofdstukken is besproken. De LVS-toetsen Spelling 3.0 voor groep 3 uit het Cito Volgsysteem primair en speciaal onderwijs vormen een hulpmiddel om vast te stellen in hoeverre leerlingen kunnen spellen. De toetsen kunnen, in samenhang met de toetsen Spelling 3.0 voor de hogere leerjaren, worden gebruikt om de spellingvaardigheid 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 Leerstoflijnen begrippenlijst en taalverzorging en recente publicaties over spelling. Deze bronnen vormden een adequate basis voor de domeinbeschrijving van de toetsen Spelling. In de domeinbeschrijving legden we uit welke aspecten en principes een rol spelen bij het leren spellen 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 spelling kan worden opgevat als een unidimensionaal continuüm. Verder werd in hoofdstuk 2 het gehanteerde meetmodel beschreven, dat gebaseerd is op de itemresponstheorie.

In aansluiting op deze theoretische uitgangspunten is in hoofdstuk 3 de domeinbeschrijving voor de toetsen Spelling verder uitgewerkt en verantwoord in de vorm van een aantal spellingcategorieën. Hier ligt een uitgebreide methodenanalyse aan ten grondslag. De constructie van de opgaven is afgeleid van deze domeinindeling en de definitieve selectie van opgaven in de toetsen is gebaseerd op een gewenste verdeling van opgaven binnen en over de verschillende spellingcategorieën. Ook is in dit hoofdstuk verslag gedaan van de itemconstructie, de opzet van de proeftoetsingen en de normeringsonderzoeken en de samenstelling van de definitieve toetsen. Ten slotte bevat hoofdstuk 3 een beschrijving van enkele psychometrische kenmerken.

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. 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. Dit betekent dat alle toetsitems succesvol konden worden geschaald en dat het functioneren van leerlingen op de toetsen terug te voeren is op één unidimensionaal concept: spellingvaardigheid. In paragraaf 4.3.2 werd voorts aangetoond dat de normeringssteekproeven op basis van de variabelen regio, urbanisatiegraad, schooltype en sekse een zeer 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 schaalscores de grenzen van de niveau-indelingen samenvallen.

In hoofdstuk 5 stond de betrouwbaarheid van de toets centraal. De betrouwbaarheidscoëfficiënt van de toetsen is met 0,86 (toetsen M3E3 en E3) en 0,92 (toets M3) 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 gehanteerde uitgangspunten en bronnen, de analyse van lesmethoden en domeinbeschrijving, de inhoudelijke verantwoording van de spellingcategorieën, constructieprocedures en itemselectie op basis van empirisch onderzoek (zie hierboven). Een eerste belangrijke grondslag voor de *begripsvaliditeit* is te vinden in het unidimensionale

karakter van de toets, zoals dat in hoofdstuk 4 is aangetoond. Uit de resultaten van de betreffende kalibratieanalyses is al af te leiden dat de kwaliteit van de items hoog is. Dit wordt bevestigd door de ‘klassieke’ itemparameters. DIF-onderzoek toont daarnaast aan dat er bij slechts enkele items (in de toets M3E3) sprake is van differentieel functioneren met betrekking tot sekse.
In hoofdstuk 6 bespraken we ook de soortgenootvaliditeit. Als belangrijke schakel in de bewijsvoering werd de hoge correlatie van de LVS-toetsen Spelling 3.0 met de LVS-toetsen Spelling van de tweede generatie opgevoerd. Deze laatste toetsen werden eerder door de COTAN op begripsvaliditeit positief beoordeeld. Op basis hiervan is de soortgenootvaliditeit van de nieuwe toets hoog te noemen. De correlaties met andere toetsen op het gebied van leervorderingen bleken lager dan de correlatie tussen de toetsen Spelling 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 Spelling 3.0 voor groep 3 begripsvalide instrumenten zijn om de spellingvaardigheid te beschrijven en te volgen.

8 Literatuur

Beek, A. van der & Paus, H. (2011). *Leerstoflijnen begrippenlijst en taalverzorging beschreven. Uitwerking van het referentiekader Nederlandse taal voor het domein begrippenlijst en taalverzorging op de basisschool.* Enschede: SLO.

Bloom, H.S., Bos, J.M., & Lee, S. (1999). Using cluster random assignment to measure program impacts. *Evaluation Review, 23*, 445-469.

Bloom, H.S., Richburg-Hayes, L., & Black, A.R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis, 29*, 30-59.

Bon, W.H.J. van (1993). *Spellingproblemen: Theorie en praktijk*. Rotterdam: Lemniscaat.

Bonset, H., & M. Hoogeveen (2009). *Spelling in het basisonderwijs. Een inventarisatie van empirisch onderzoek*. Enschede: SLO

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

Cito (2014). *Cito Volgsysteem primair en speciaal onderwijs. Spelling 3.0 Groep 3*. Arnhem: Cito.

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

Eggen, T.J.H.M. (1993). Itemresponstheorie en onvolledige gegevens. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 239-284). Arnhem: Cito.

Engelen, R.J.H. en Eggen, T.J.H.M., (1993). Equivaleren. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 309-348). Arnhem: Cito.

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: Expertgroep doorlopende leerlijnen Taal en Rekenen.

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

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

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

Gijsel, M., Scheltinga, F., Druenen, M. van & Verhoeven, L. (2011a). *Protocol Leesproblemen en Dyslexie voor groep 3*. Nijmegen: Expertisecentrum Nederlands.

Gijsel, M., Scheltinga, F., Druenen, M. van & Verhoeven, L. (2011b). *Protocol Leesproblemen en Dyslexie voor groep 4*. Nijmegen: Expertisecentrum Nederlands.

Glas, C.A.W. & N.D. Verhelst (1993). Een overzicht van itemresponsmodellen. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 179-238). Arnhem: Cito.

Hedges, L.V., & Hedberg, E.C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis, 29*, 60-87.

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.

Huizenga, H. (2010). *Taal & didactiek. Spelling* (4e herziene druk). Groningen: Noordhoff Uitgevers.

Keuning, J. (2011). *Normeren op school met Cito dataretour*. Arnhem: Cito.

Keuning, J., Boxtel, H. van, Lansink, N., Visser, J., Weekers, A. & Engelen, R. (2014). *Actualiteit en kwaliteit van normen. Een werkwijze voor het normeren van een leerlingvolgsysteem.* Arnhem: Cito.

Kleijnen, R. (1997). *Strategieën van zwakke lezers en spellers in het voortgezet onderwijs. Dissertatie Vrije Universiteit*. Lisse: Swets en Zeitlinger.

Kleijnen, R. (2004). *Hardnekkige spellingfouten. Een taalkundige analyse*. Lisse: Harcourt Book publishers.

Kuhlemeier, H., Til, A. van, Hemker, B., Klijn, W. de & Feenstra, H. (2013). *Balans van de schrijfvaardigheid in het basis- en speciaal basisonderwijs 2. Uitkomsten van de peiling in 2009 in groep 5, groep 8 en de eindgroep van het SBO.* PPON-reeks nummer 53. Arnhem: Cito.

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

Ministerie van Onderwijs, Cultuur en Wetenschappen (2006). *Kerndoelen primair onderwijs.* Den Haag: MinOCW.

Nederlandse Taalunie (2009). *Technische handleiding. Regels voor de officiële spelling van het Nederlands*. Geraadpleegd op 19 juni 2014 via http://taalunieversum.org/inhoud/spelling-meer-hulpmiddelen/technische-handleiding

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.

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

Schijf, G.M. (2009). *Lees- en spellingvaardigheden van brugklassers* (proefschrift). Amsterdam: SCO-Kohnstamm Instituut, Universiteit van Amsterdam.

Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33*, 62-87.

Schryver, J. de & A. Neijt (2005). *Handboek Spelling* (5e herziene druk). Mechelen: Wolters Plantyn.

Snijders, T.A.B. & Bosker, R.J. (1993). *Standard errors and sample sizes for two-level research*. Journal of Educational Statistics, 18, 237-260.

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

Til, A. van, Weerden, J. van, Hemker, B. & Keune, K. (2014). *Balans van de taalverzorging en grammatica in het basis- en speciaal basisonderwijs. Uitkomsten van de peiling in 2009 in jaargroep 5, jaargroep 8 en de eindgroep van het SBO*. PPON-reeks nummer 55. Arnhem: Cito.

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

Verhelst, N.D. & F.G.M. Kleintjes (1993). Toepassingen van itemresponsetheorie. In: T.J.H.M. Eggen en P.F. Sanders (Red.). *Psychometrie in de praktijk*. Arnhem: Cito.

Verhelst, N.D., H.H.F.M. Verstralen & T.H.J.M. Eggen (1991). Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model. *Measurement and Research Department Reports 91-10*. Arnhem: Cito.

Woordenlijst van de Nederlandse Taal (2005). Samengesteld door Instituut voor Nederlandse Lexicologie in opdracht van de Nederlandse Taalunie. Den Haag: SDU.

Bijlagen

Bijlage 1 Klassieke en IRT-indices van de opgaven in toetsen Spelling 3.0 groep 3

Toets M3

Volgnr	P-Val	RIT	Beta	Info
1.1	0,759	0,555	-0,327	1,985
1.2	0,741	0,666	-0,235	3,616
1.3	0,761	0,555	-0,33	1,977
1.4	0,922	0,244	-1,286	0,268
1.5	0,833	0,317	-0,812	0,5
1.6	0,805	0,451	-0,511	1,118
1.7	0,682	0,482	-0,229	1,477
1.8	0,86	0,301	-0,927	0,438
1.9	0,773	0,464	-0,429	1,232
1.10	0,771	0,464	-0,425	1,237
1.11	0,746	0,471	-0,366	1,315
1.12	0,655	0,366	-0,265	0,778
1.13	0,785	0,339	-0,639	0,597
1.14	0,707	0,668	-0,181	3,844
1.15	0,915	0,252	-1,235	0,29
1.16	0,558	0,366	-0,024	0,842
1.17	0,801	0,54	-0,416	1,767
1.18	0,806	0,45	-0,514	1,113
1.19	0,867	0,296	-0,96	0,421
1.20	0,919	0,247	-1,265	0,277
2.1	0,92	0,35	-0,924	0,571
2.2	0,72	0,477	-0,309	1,387
2.3	0,836	0,433	-0,6	0,992
2.4	0,758	0,618	-0,286	2,736
2.5	0,874	0,404	-0,726	0,817
2.6	0,86	0,567	-0,504	1,929
2.7	0,826	0,44	-0,569	1,036
2.8	0,793	0,336	-0,664	0,583
2.9	0,817	0,444	-0,546	1,069
2.10	0,777	0,613	-0,321	2,61
2.11	0,811	0,535	-0,439	1,71
2.12	0,763	0,467	-0,405	1,264
2.13	0,887	0,392	-0,774	0,754
2.14	0,704	0,479	-0,273	1,429
2.15	0,72	0,477	-0,309	1,388
2.16	0,762	0,617	-0,294	2,708
2.17	0,784	0,547	-0,378	1,86
2.18	0,44	0,35	0,254	0,841
2.19	0,913	0,36	-0,892	0,608
2.20	0,723	0,476	-0,315	1,38

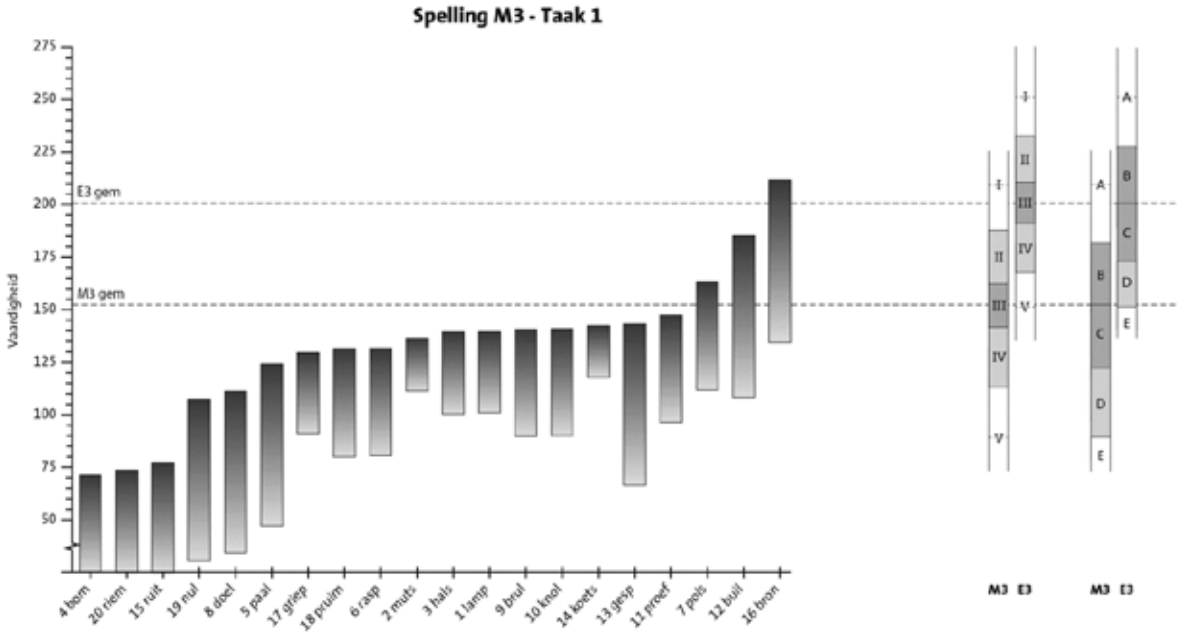
Toets M3E3

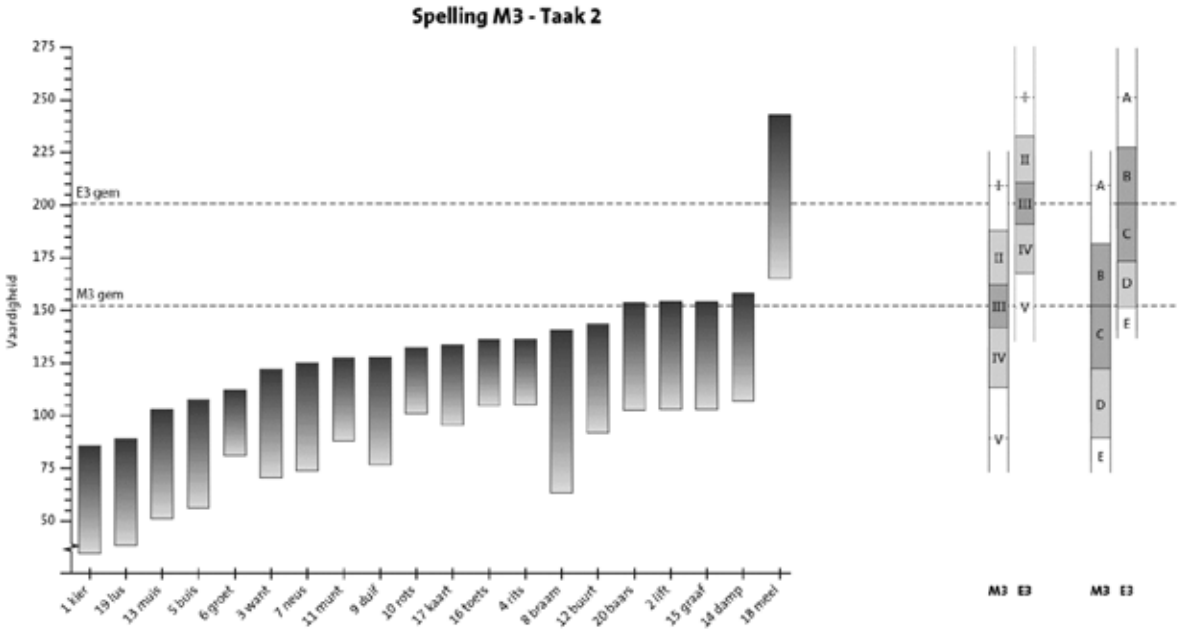
Volgnr	P-Val	RIT	Beta	Info
1.1	0,931	0,285	-0,459	0,53
1.2	0,943	0,265	-0,538	0,444
1.3	0,934	0,359	-0,299	0,842
1.4	0,849	0,359	-0,134	0,996
1.5	0,707	0,401	0,202	1,534
1.6	0,851	0,443	-0,022	1,602
1.7	0,868	0,432	-0,066	1,466
1.8	0,927	0,29	-0,438	0,555
1.9	0,815	0,46	0,06	1,86
1.10	0,663	0,403	0,285	1,639
1.11	0,948	0,333	-0,372	0,688
1.12	0,894	0,327	-0,286	0,758
1.13	0,945	0,261	-0,55	0,432
1.14	0,895	0,226	-0,606	0,357
1.15	0,915	0,305	-0,377	0,632
1.16	0,829	0,454	0,029	1,761
1.17	0,88	0,338	-0,235	0,835
1.18	0,729	0,398	0,158	1,471
1.19	0,815	0,376	-0,037	1,156
1.20	0,934	0,28	-0,478	0,509
2.1	0,94	0,348	-0,331	0,771
2.2	0,903	0,318	-0,323	0,706
2.3	0,901	0,32	-0,316	0,715
2.4	0,877	0,341	-0,223	0,854
2.5	0,885	0,335	-0,252	0,809
2.6	0,788	0,386	0,03	1,268
2.7	0,901	0,403	-0,168	1,172
2.8	0,894	0,326	-0,289	0,755
2.9	0,924	0,201	-0,789	0,271
2.10	0,907	0,397	-0,188	1,116
2.11	0,797	0,466	0,098	1,977
2.12	0,89	0,33	-0,27	0,782
2.13	0,857	0,355	-0,157	0,959
2.14	0,702	0,401	0,211	1,546
2.15	0,949	0,33	-0,379	0,673
2.16	0,943	0,343	-0,344	0,743
2.17	0,941	0,346	-0,335	0,763
2.18	0,864	0,434	-0,057	1,493
2.19	0,757	0,393	0,1	1,382
2.20	0,924	0,2	-0,793	0,269

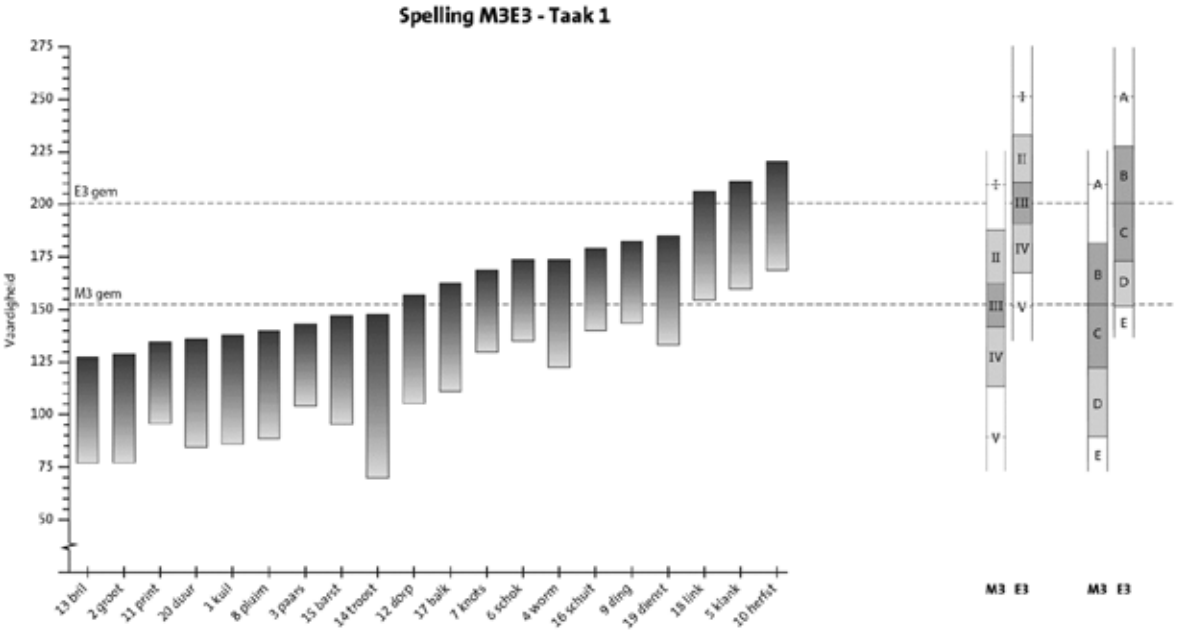
Toets E3

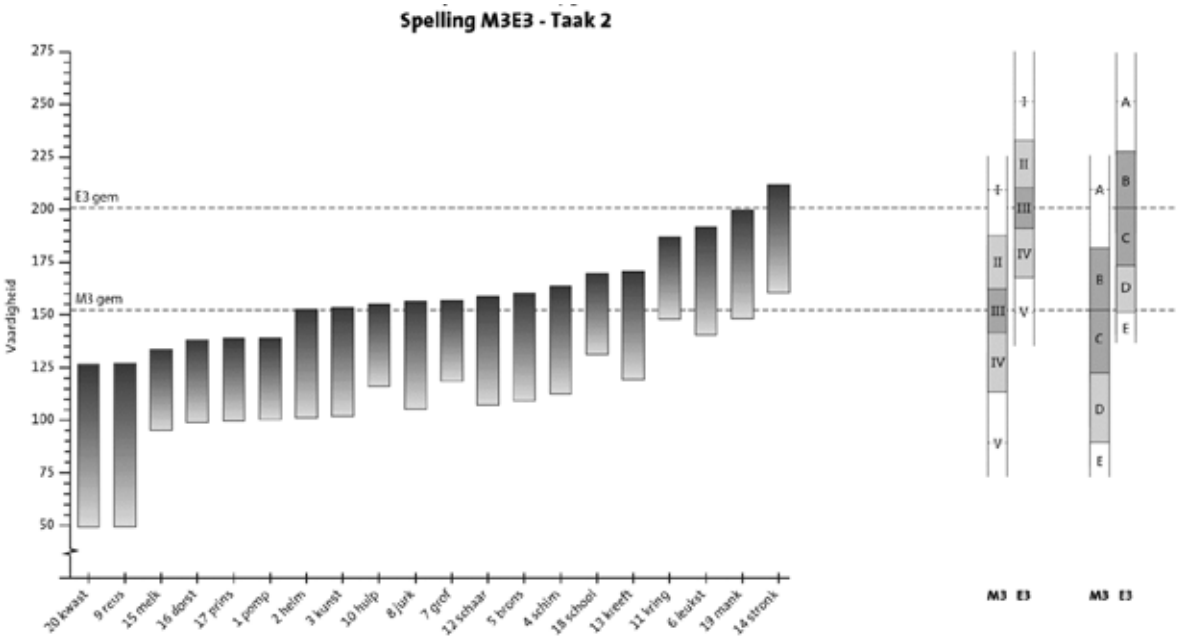
Volgnr	P-Val	RIT	Beta	Info
1.1	0,875	0,338	-0,218	1,116
1.2	0,695	0,492	0,279	0,805
1.3	0,725	0,407	0,167	1,678
1.4	0,839	0,446	0,006	1,056
1.5	0,813	0,378	-0,033	1,002
1.6	0,456	0,306	0,664	1,132
1.7	0,836	0,366	-0,094	2,493
1.8	0,865	0,347	-0,183	0,652
1.9	0,811	0,379	-0,027	1,435
1.10	0,832	0,368	-0,082	1,047
1.11	0,855	0,354	-0,149	0,857
1.12	0,861	0,249	-0,439	1,065
1.13	0,879	0,237	-0,524	0,926
1.14	0,83	0,369	-0,077	2,121
1.15	0,865	0,347	-0,183	0,463
1.16	0,634	0,417	0,337	0,877
1.17	0,797	0,279	-0,193	1,321
1.18	0,838	0,365	-0,1	1,563
1.19	0,876	0,338	-0,221	0,917
1.20	0,774	0,393	0,062	1,038
2.1	0,824	0,372	-0,062	0,408
2.2	0,877	0,239	-0,51	0,862
2.3	0,644	0,416	0,32	0,887
2.4	0,837	0,365	-0,098	1,484
2.5	0,848	0,358	-0,13	1,69
2.6	0,547	0,313	0,46	1,164
2.7	0,82	0,374	-0,052	0,887
2.8	0,769	0,289	-0,102	1,062
2.9	0,74	0,403	0,135	0,917
2.10	0,839	0,364	-0,103	1,173
2.11	0,841	0,363	-0,108	1,081
2.12	0,835	0,366	-0,092	0,97
2.13	0,863	0,348	-0,177	0,448
2.14	0,773	0,476	0,145	0,4
2.15	0,856	0,252	-0,414	1,091
2.16	0,873	0,341	-0,208	0,917
2.17	0,774	0,393	0,062	1,695
2.18	0,696	0,411	0,224	0,596
2.19	0,865	0,347	-0,183	1,052
2.20	0,886	0,33	-0,255	1,321

Bijlage 2 Moeilijkheid van opgaven per taak in Spelling 3.0 groep 3

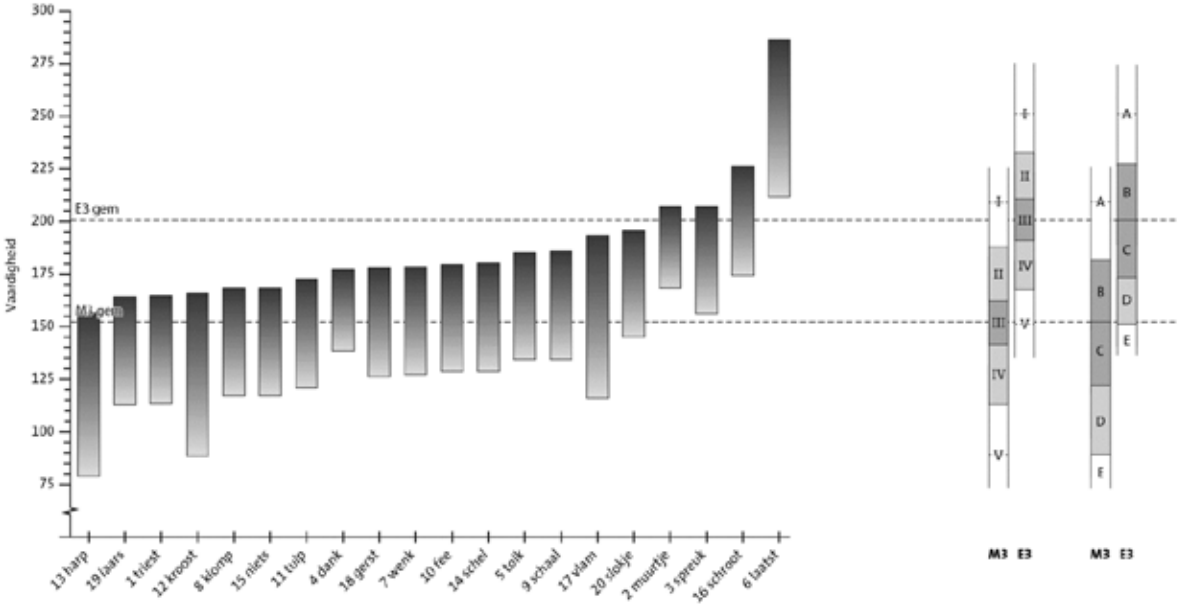




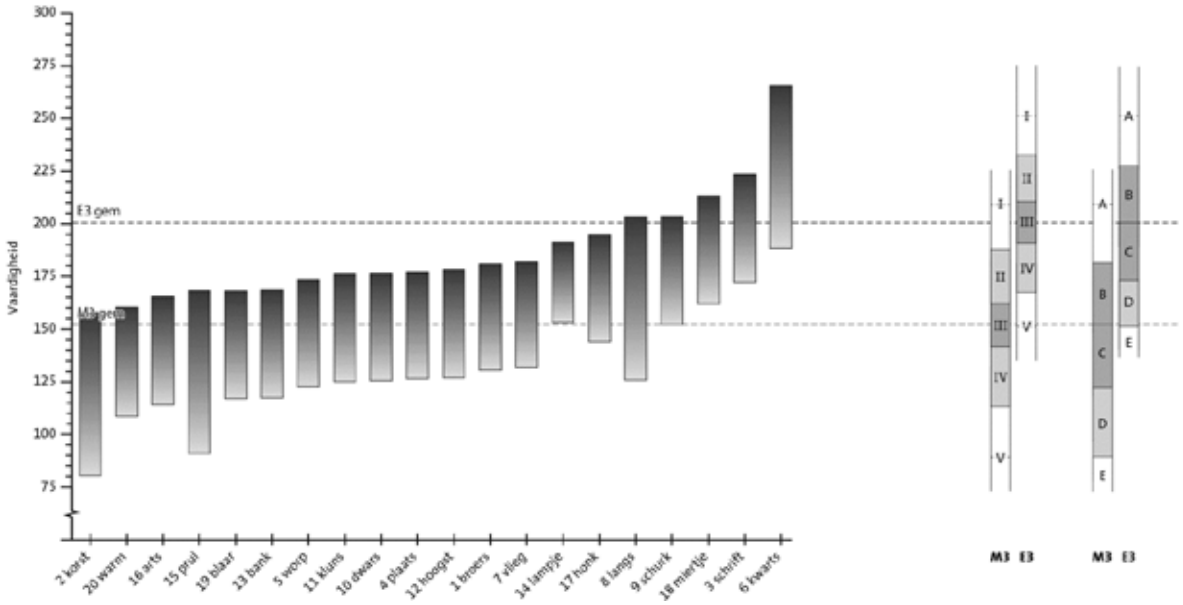

Spelling M3E3 - Taak 1
Vaardigheid
E3 gem
M3 gem
13 bril
2 groet
11 print
20 duur
1 kuil
8 plaim
3 paars
15 barst
14 troost
12 dorp
17 balk
7 knots
6 schok
4 worm
16 schuit
9 ding
19 dienst
18 link
5 klank
10 herfst
M3 E3
M3 E3


Spelling M3E3 - Taak 2
Vaardigheid
E3 gem
M3 gem
20 kwast
9 reus
15 melk
16 dorst
17 prins
1 pomp
2 heim
3 kunst
10 hulp
8 jurk
7 grof
12 schaar
5 brons
4 schim
18 school
13 kreeft
11 kring
6 keukst
19 mank
14 stronk
M3 E3
M3 E3

Spelling E3 - Taak 1



Spelling E3 - Taak 2



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