# Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8

Marieke Tomesen, Jasper Wouda, Irene Krämer en Linda Horsels

# Wetenschappelijke verantwoording van de LVS-toetsen

## Spelling 3.0 groep 8

Marieke Tomesen
Jasper Wouda
Irene Krämer
Linda Horsels

Cito
Arnhem, 2019

# Inhoud

**Bijlagen 81**

# 1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de papieren LVS-toetsen Spelling 3.0 voor groep 8. Het toetspakket Spelling 3.0 voor groep 8 bevat twee toetsen: een toets voor het spellen van niet-werkwoorden (Spelling niet-werkwoorden B8/M8) en een toets voor het spellen van werkwoorden (Spelling werkwoorden B8/M8).
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 basisonderwijs. Ze zijn ook geschikt voor toepassing in het speciaal basisonderwijs en het speciaal onderwijs cluster 1, cluster 2 (TOS-leerlingen) en cluster 4. De toetsen Spelling 3.0 betreffen papieren en digitale toetsen voor alle leerjaren.

Reeds eerder zijn de wetenschappelijke verantwoordingen voor de papieren toetsen Spelling 3.0 groep 3 t/m 7 uitgebracht, evenals de wetenschappelijke verantwoordingen voor de digitale toetsen groep 3, 4 en 5. Te zijner tijd zullen ook de wetenschappelijke verantwoordingen voor de digitale toetsen groep 6 t/m 8 worden uitgebracht.

Deze verantwoording biedt tezamen met de inhoud van het toetspakket Spelling 3.0 voor groep 8 alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit van de betreffende meetinstrumenten. Het genoemde materiaal maakt een beoordeling van de toetsen Spelling 3.0 groep 8 mogelijk op de volgende aspecten:
- Uitgangspunten van de toetsconstructie;
- De kwaliteit van het toetsmateriaal;
- De kwaliteit van de handleiding;
- Normen; N.B.: De toetsen voor groep 8 zijn op twee momenten in het leerjaar af te nemen: half oktober/half november (B8-moment) en half januari/half februari (M8-moment). Deze wetenschappelijke verantwoording gaat alleen in op de nornering van M8. In 2020 zal een addendum over B8 verschijnen.
- 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 8. De kwaliteit van het toetsmateriaal en de handleiding is te bepalen door kennis te nemen van de inhoud van het toetspakket.

# 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.
Het toetspakket Spelling 3.0 voor groep 8 omvat zowel toetsen voor de spelling van niet-werkwoorden als toetsen werkwoordspelling. De toetsen zijn bedoeld om vast te stellen hoe goed een leerling (onveranderlijke woorden cq. werkwoordsvormen) 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).

## 2.2 Doelgroep

De toetsen Spelling 3.0 voor groep 8 van het Cito Volgsysteem primair en speciaal onderwijs zijn bestemd voor leerlingen in groep 8 van het basisonderwijs. De toetsen zijn ook geschikt voor leerlingen in het speciaal basisonderwijs en het speciaal onderwijs cluster 1, cluster 2 (TOS-leerlingen) en cluster 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 8 zijn zowel voor ‘begin leerjaar’ (half oktober/half november) als voor ‘midden leerjaar’ (half januari/half februari) populatieparameters bepaald[1]. Voor beide onderdelen, werkwoordspelling en niet-werkwoordspelling, is er één toets. De leerkracht of school bepaalt zelf op welk van deze momenten de toetsen worden afgenomen. De toetsen kunnen desgewenst ook op een ander moment in het schooljaar worden afgenomen, maar dat maakt het moeilijker om uitspraken te doen over het niveau van de leerling ten opzichte van andere leerlingen in Nederland.
De toetsen Spelling 3.0 voor groep 8 kunnen ook gebruikt worden voor leerlingen in lagere leerjaren die werken op het niveau van groep 8. In de handleiding is toegelicht hoe dit toetsen op maat, met behulp van vaardigheidsscores, in zijn werk gaat. Ook is het mogelijk dat leerlingen in groep 8 een toets van een lager niveau maken. Voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften zijn in de handleiding extra aanwijzingen opgenomen. Voor deze leerlingen zijn alternatieve rapportageformulieren ontwikkeld.

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 gehoor-problemen. 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 in staat geacht om aan de hand van de aanwijzingen in de handleiding een gestandaardiseerde en ongestoorde toetsafname te realiseren.

[1] Over B8 volgt een addendum op deze wetenschappelijke verantwoording.

## 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 hoofdstuk 4 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.

*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 |

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

**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 LVS-toetsen 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 het spellen van niet-werkwoorden, respectievelijk het spellen van werkwoorden, 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 van dezelfde vaardigheid, 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). Er is een aparte schaal voor Spelling niet-werkwoorden en een schaal voor Spelling werkwoorden. Beide vaardigheden verschillen namelijk dusdanig van elkaar, dat ze niet op één schaal passen.
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? Onderstaande geldt voor Spelling niet-werkwoorden en Spelling werkwoorden afzonderlijk.
Globaal kunnen we de toetsresultaten van leerlingen (of groepen leerlingen) op twee manieren interpreteren:

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

Bij beide vergelijkingen maken we gebruik van het feit dat de toetsresultaten door toepassing van het IRT-model (OPLM) in de vorm van vaardigheidsscores afgebeeld kunnen worden op dezelfde vaardigheidsschaal. Dat geldt voor zowel individuele leerlingen als voor (gemiddelde) 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 ingedeeld in vaardigheidsniveaus (I-V, A-E), op basis van de data die verzameld zijn in het kader van het normeringsonderzoek (zie hiervoor hoofdstuk 4 van deze wetenschappelijke verantwoording). Vaardigheidsniveau I, bijvoorbeeld, bevat de 20% hoogst scorende leerlingen. Door de vaardigheidsscore van een leerling te vergelijken met deze vaardigheids-niveaus (die zijn afgebakend door percentielpunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk als "Koen heeft op afnamemoment medio leerjaar 8 vaardigheidsniveau IV behaald". Voor de leerkracht (en voor Koen en zijn ouders) bevat deze uitspraak waardevolle informatie.
De leerkracht kan op basis hiervan bijvoorbeeld besluiten om Koen extra lesstof aan te bieden.

Ad b.
Voor het vergelijken (‘volgen’) van een leerling op twee verschillende tijdstippen komen twee methodes in aanmerking. Bij de eerste methode worden de **vaardigheidsniveaus** op de twee tijdstippen vergeleken: “op tijdstip E7 had Koen vaardigheidsniveau III en op tijdstip M8 is het vaardigheidsniveau IV”. Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 354, bijvoorbeeld, op tijdstip E7 en vaardigheidsscore 359 op tijdstip M8. 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 bij ad a. als bij ad b., dienen uitspraken over leerlingen te worden gerelativeerd. In de handleiding is meer informatie te vinden over de wijze waarop de gebruiker dit kan doen. Hieronder gaan we vooral in op het belang van de (on)betrouwbaarheid van de afgenomen toetsen hierbij.
Voor elke toets geldt dat de vaardigheidsscore die bij een toetsresultaat van een leerling hoort, behept is met een meetfout. Als we rekening houden met die meetfout dan zou het best zo kunnen zijn dat Koen vaardigheidsniveau III heeft behaald op het tweede tijdstip en niet vaardigheidsniveau IV. Of dat we moeten concluderen dat het verschil in de prestaties op de twee tijdstippen statistisch niet significant is: Koen is voor- noch achteruitgegaan. In alle gevallen speelt het betrouwbaarheidsinterval (BI) rondom de vaardigheidsscore een belangrijke rol. Dat geldt ook voor de indeling in vaardigheidsniveaus. Hoe het BI doorwerkt in de indeling in vaardigheidsniveaus en de verdere gevolgen daarvan wordt beschreven in hoofdstuk 5. Op deze plaats beperken we ons tot de vraag of we de uitspraak kunnen doen dat een leerling of groep (werkelijk) ‘gegroeid’ is. De eenvoudigste manier is om te kijken of de BI’s voor de twee tijdstippen overlappen. Als deze twee BI’s niet overlappen dan is er sprake van een significant verschil in vaardigheid tussen beide tijdstippen. Overlappen ze wel dan is er geen verschil in vaardigheid. Deze eenvoudige manier van vergelijken kan de leerkracht zelf uitvoeren en wordt ook in de handleiding beschreven. We geven hier een voorbeeld. Bij de afname Spelling niet-werkwoorden E7 behaalde Samira een vaardigheidsscore van 349 met een 67% betrouwbaarheidsinterval van 341-357. Bij de afname M8 behaalde Samira een vaardigheidsscore van 367; het bijbehorende betrouwbaarheidsinterval daarbij is 359-375. Aangezien de betrouwbaarheidsintervallen niet overlappen kunnen we zeggen dat Samira’s vaardigheid is toegenomen.

Conclusie
De vaardigheidsgroei voor Spelling (zowel niet-werkwoorden als werkwoorden) voltrekt zich langzaam in de tijd. De verschillen tussen vaardigheidsscores op achtereenvolgende meettijdstippen zijn betrekkelijk klein. Bovendien is er sprake van meetfouten. De verschillen in vaardigheidsgroei moeten tegen de achtergrond van die meetfouten worden geïnterpreteerd. Dit betekent dat men weliswaar uitspraken kan doen over de vaardigheidsgroei van een leerling, maar dat deze uitspraken met voorzichtigheid dienen te worden gehanteerd. Dit geldt ook wanneer men de progressie van een leerling volgt in termen van vaardigheidsniveaus of een vergelijking maakt met andere leerlingen in termen van vaardigheidsniveaus. Want ook bij de indeling in vaardigheidsniveaus speelt de nauwkeurigheid van de toets een rol. Hoe de leerkracht hier in de praktijk mee om moet gaan, wordt toegelicht in de handleiding voor de leerkracht.

**'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 leerlingantwoorden 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 niet-werkwoorden beperkt is (er zijn veel categorieën in de toetsen Spelling niet-werkwoorden en de toets mag niet te lang worden), kan niet worden uitgesloten dat de leerling bij toeval 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 niet-werkwoorden. 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). Voor Spelling werkwoorden zijn geen controledictees opgenomen; in de toetsen Spelling werkwoorden komen relatief meer opgaven (gemiddeld 5) per categorie voor dan in de toetsen Spelling niet-werkwoorden waardoor een betere inschatting te maken is of de leerling een categorie al dan niet beheerst.
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 het bieden van een handreiking aan leerkrachten die gericht extra ondersteuning willen geven aan leerlingen die moeite hebben met het correct spellen van bepaalde woorden.

## 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 taal-vaardigheden 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 uitgave 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 3 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 Nederlandse spelling hulptekens om de klank van (groepen) letters duidelijker weer te 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 uitspraak-nuances die ontstaan door persoonsgebonden of streekgebonden verschillen of door klanken in de omgeving van het woord genegeerd. Bijvoorbeeld: de ‘z’ in ‘zet’ klinkt in grote delen van Nederland als de stemloze /s/ , maar wordt toch als z geschreven. Waar sprekers de /z/ wel mét stem uitspreken, klinkt toch de /z/ in 'ik **z**et' als een /s/ in tegenstelling tot de /z/ in '**z**et ik'. Ook deze uitspraaknuances worden genegeerd. 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. 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 16$^{e}$ 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 de 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 beschrijft 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. Dit criterium is alleen van toepassing op de spelling van niet-werkwoorden, niet op de werkwoordspelling.
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 te moeten opzoeken. Het gaat dan om woorden als: museum, enigszins, directie, chauffeur. In het onderwijs in werkwoordspelling speelt het moeilijkheidscriterium ook een rol, zij het een andere: werkwoordsvomen waarin de spelling overeenkomt met regels voor het spellen van niet-werkwoorden krijgen weinig aandacht. Dit betreft bijvoorbeeld een vorm als loopt, waarbij nauwelijks fouten worden gemaakt.
Ad 3. Tot slot wordt voor de ordening van de leerstof verder uitgegaan van een indeling in

spellingcategorieën, 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. Dit geldt zowel voor werkwoordspelling als voor spelling van niet-werkwoorden. 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 onderdelen van het woord, ook wel woorddelen genoemd. Doordat deze delen altijd hetzelfde worden geschreven, komt de leerling tot een correcte spelling. Het volstaat om naar het woord zelf te kijken. Woorden als ondiep (voorvoegsel), 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. Wel beginnen alle methoden met de klankzuivere woorden en wordt er in de meeste methoden in groep 6 een begin gemaakt met de werkwoordspelling (morfologische spelling op syntactische basis). Categorieën die op een bepaald moment nieuw worden aangeboden, worden in de daaropvolgende leerjaren steeds herhaald.

**De bijzondere positie van werkwoordspelling**

De werkwoordspelling neemt binnen het spellingonderwijs, en dus ook binnen de toetsen Spelling 3.0, een bijzondere positie in. In veel spellingmethoden gaat pas in groep 7 en 8 veel aandacht uit naar werkwoordspelling. Een aantal methoden introduceert delen van de werkwoordspelling al eerder.
Het gaat bij werkwoordspelling altijd om morfologische spelling op syntactische basis, waarbij dus de juiste spelling van het woord afhankelijk is van de relatie tot andere delen van de zin of tekst. De regels voor werkwoordspelling zijn alléén bij werkwoorden van toepassing. Dit maakt dat de spellingstrategieën die leerlingen hanteren voor de algemene spelling, bij werkwoorden relatief vaak tot een incorrecte spelling leiden. Bij niet-werkwoorden zagen we al dat de fonologische strategie tot spelfouten kan leiden: *ijselijk* kan

met een fonologische strategie gespeld worden als *eiselek*. Op dezelfde manier kan een fonologische strategie leiden tot het spellen van *vind* als *vint*, of van *leefden* als *leevde*. Bij werkwoordspelling is daarnaast ook de directe strategie of de woordbeeldstrategie slechts heel beperkt toepasbaar. Terwijl wij het woord *ijselijk* goed kunnen spellen, omdat wij het nu eenmaal altijd zo gespeld zien, blijkt dat een woord dat klinkt als *wachten* op allerlei verschillende manieren gespeld wordt: *wachtte*, *wachten* en *wachtten*. Alleen kijken naar het het woord zelf, zoals bij de spelling voor niet-werkwoorden, is dus niet voldoende: getal, persoon en tijd hebben allemaal invloed op de spelling. Werkwoorden correct spellen kan alleen op basis van een - vrij gecompliceerde - regelstrategie, vaak gecombineerd met een analogiestrategie. Een voorbeeld van een regelstrategie is: in de 2e en 3e persoon enkelvoud tegenwoordige tijd krijg je stam+t, een voorbeeld van een analogiestrategie is: je schrijft 'hij loopt' met een -t, dan schrijf je dus ook 'hij vindt' met een -t.

De moeilijkheid van spelling is er niet zozeer in gelegen, dat de speller één correcte spellingstrategie moet kiezen. Het is namelijk zo, dat iedere speller, zowel de beginnende als de gevorderde, meerdere spellingstrategieën tegelijkertijd inzet. Het is juist de kunst om, daar waar de verschillende strategieën een verschillende uitkomst opleveren, aan de juiste strategie voorrang te verlenen.
Spellers zetten gelijktijdig meerdere spellingstrategieën in en nemen voortdurend beslissingen welke strategie voorrang krijgt (Huizenga, 2010). Dat de woordbeeldstrategie meestal actief is, blijkt bijvoorbeeld daaruit, dat geoefende spellers direct de vorm 'beloovd' herkennen als fout: deze vorm zie je immers nooit. Met behulp van dit voorbeeld kunnen we ook meteen duidelijk maken, waarom de woordbeeldstrategie niet geheel betrouwbaar is. De fout in 'zij had het me nog zo belooft' onstnapt veel vaker aan de aandacht van de speller dan de fout 'beloovd'. De woordbeeldstrategie, die een goed resultaat oplevert bij 'beloovd', leidt tot het verkeerde resultaat in 'zij had belooft', waar we een regelstrategie moeten toepassen. Door het voorkomen van gelijkklinkende vormen is het bij de Nederlandse werkwoordspelling bijzonder lastig om te kiezen voor de juiste strategie en, binnen de regelstrategie, voor de juiste regel (Huizenga, 2010; Verhaert & Sandra, 2016).
Kortom, het beschikbaar zijn van meerdere gelijkklinkende vormen, het gebruik van ontoereikende strategieën én het verkeerd toepassen van een regel binnen de regelstrategie kunnen allemaal tot het verkeerde resultaat leiden bij het spellen van een werkwoord. Voeg daarbij nog het hanteren van nieuwe concepten in de lessen ('voltooid deelwoord', 'dt', 'stam') en het toepassen van basaal grammaticaal inzicht, en het is duidelijk waarom het onderwijs bijzondere aandacht aan de werkwoordspelling besteedt. Deze bijzondere aandacht vormt, met het feit dat de regels voor werkwoordspelling zeer verschillend zijn van de andere Nederlandse spellingregels, de reden voor het opnemen van aparte toetsen voor werkwoordspelling in de toetsen Spelling 3.0.

**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 streefniveau. 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 taalverzorging 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 opgavenbanken. 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 gebruikgemaakt van de opgavenbanken Spelling niet-werkwoorden en Spelling werkwoorden. Voor andere vakgebieden van 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 niet-werkwoorden alsook de vaardigheid werkwoordspelling 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 de vaardigheid uit, waarbij een groter getal wijst op een grotere vaardigheid. Het doel van de meetprocedure – het afnemen van een toets – is de plaats van de leerling op dit continuüm zo nauwkeurig mogelijk te bepalen. De uitkomst van de meetprocedure bestaat strikt genomen uit twee grootheden: de eerste is de schatting van de plaats van de leerling op het vaardigheidscontinuüm. De tweede 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 niet-werkwoorden het spellen van niet-werkwoorden meten en dat alle items in de opgavenbank Spelling werkwoorden werkwoordspelling 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 zogenoemde 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 zogenoemde gekalibreerde itembank. In dat proces worden de items die niet passen bij de verzameling uit de collectie verwijderd. De opgavenbank bevat voor elk item niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen, en de statistische zekerheid dat alle items dezelfde vaardigheid aanspreken. Dit houdt onder meer het volgende in:

1 In principe kan met een willekeurige selectie items uit de bank de vaardigheid worden gemeten bij een willekeurige leerling. In principe, want een willekeurige toets die uit de itembank wordt getrokken, zal in de praktijk meestal niet voldoen omdat de meetresultaten (de schatting van de vaardigheid) onvoldoende nauwkeurig 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 (medio groep 7 in het geval van Spelling werkwoorden) 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 8 kan een toets maken die normaliter aan groep 7 wordt voorgelegd, en zijn vaardigheidsschatting kan behalve met de populatie van groep 8 ook vergeleken worden met de percentielen in de populatie van groep 7, met bijvoorbeeld de uitspraak: “De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 7”.

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 8 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 deze verantwoording worden de begrippen die hierboven aan de orde zijn geweest nader uitgewerkt en toegelicht voor de opgavenbanken Spelling niet-werkwoorden en Spelling werkwoorden. De verantwoording van de inhoudelijke constructie van deze opgavenbanken staat in hoofdstuk 3. In hoofdstuk 4 wordt (onder andere) de psychometrische constructie van de opgavenbanken besproken (kalibratie).

#### 2.4.2.2 Het gehanteerde meetmodel

In de normeringsonderzoeken 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 zogeheten 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 zogenoemde itemresponsfunctie. Die geeft aan hoe groot de kans is op een correct antwoord als functie van de onderliggende eigenschap of vaardigheid. Formeler: zij $X_i$ de toevalsvariabele die het antwoord op item $i$ voorstelt. $X_i$ neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord. Als symbool voor de vaardigheid wordt θ (theta) gekozen. De vaardigheid θ is niet rechtstreeks observeerbaar. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd[3]. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

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

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

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

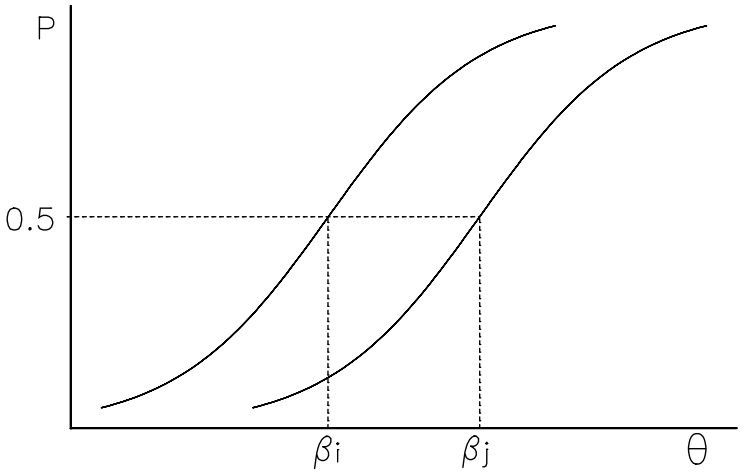
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \tag{2.2}$$

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

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

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

*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.

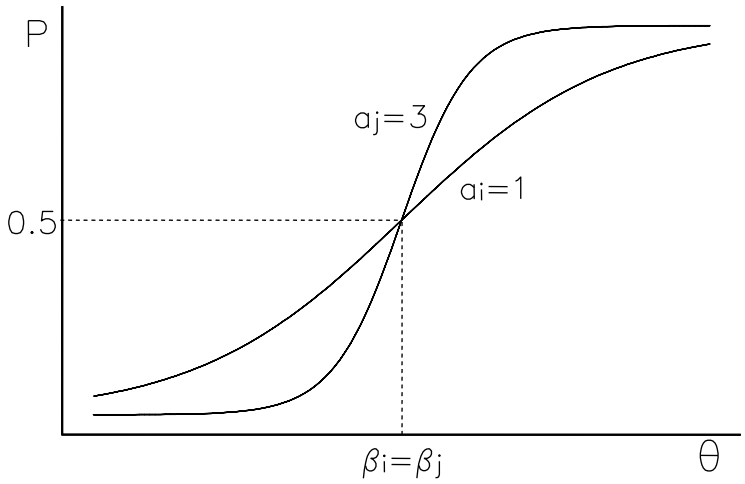
Veel van de items blijken dan ook niet beschreven te kunnen worden met het Raschmodel. Daarom is bij dit instrument gekozen voor een ander IRT-model.

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 één parameter logistisch model (One Parameter Logistic Model, afgekort: OPLM) is CML mogelijk. Dit model is, anders dan het Raschmodel, wel bestand tegen 'omwisseling' van 'proporties juist' in verschillende steekproeven (Glas & Verhelst, 1993; Eggen, 1993; Verhelst & Kleintjes, 1993). De itemresponsfunctie van het OPLM is gegeven door

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

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

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




De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas en Verstralen, 1995). Dit programma voert ook statistische toetsen uit op grond waarvan kan worden bepaald of het model de gegevens

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

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

# 3 Beschrijving van de toetsen

## 3.1 Opbouw en structuur van de toetsen

Het toetspakket Spelling 3.0 voor groep 8 uit het Cito Volsysteem primair en speciaal onderwijs bevat in totaal twee papieren toetsen: een toets Spelling niet-werkwoorden (B8/M8) en een toets Spelling werkwoorden (B8/M8).
Beide toetsen hebben een normering voor twee afnamemomenten: aan het begin van leerjaar 8 en halverwege leerjaar 8. De leerkracht of school bepaalt zelf op welk van deze momenten de toets wordt afgenomen.

*Opbouw*
De toetsen Spelling voor groep 8 bestaan uit twee taken van elk 25 opgaven. De taken van een toets dienen bij voorkeur te worden afgenomen op twee verschillende dagdelen, zodat de leerlingen geconcentreerd aan beide taken kunnen werken. De toetsen Spelling niet-werkwoorden en Spelling werkwoorden kunnen los van elkaar ingezet worden.

*Vorm*
De toetsen, zowel die voor niet-werkwoorden als die voor werkwoorden, bevatten alleen zinsdictee-opgaven. Bij een zinsdictee leest de leerkracht een zin voor en herhaalt vervolgens uit deze zin één woord. Dat woord moeten de leerlingen opschrijven. Bij de toetsen Spelling werkwoorden leest de leerkracht eerst ook nog het hele werkwoord voor en kunnen de leerlingen meelezen op het antwoordblad.

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 in de toetsen Spelling alleen actieve spelling op te nemen. Dat wil zeggen dat er dus geen meerkeuzeopgaven zijn opgenomen, wat 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. Bij de toetsen Spelling werkwoorden geldt bovendien dat de leerlingen de werkwoorden niet zelf hoeven te vervoegen. Leerlingen moeten de werkwoordsvorm die zij horen opschrijven en kunnen zich volledig concentreren op het toepassen van de regels voor het spellen van werkwoorden.
Het omschakelen naar uitsluitend dicteeopgaven levert geen problemen op voor de meetpretentie van de toetsen. Analyses van de itemgegevens van de tweede generatie toetsen lieten zien dat de verschillende opgaventypes 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 interpunctie, ondergebracht in de toetsen Taalverzorging voor groep 6 tot en met 8 van het Cito Volgsysteem voor primair en speciaal onderwijs. 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 spellingvaardigheid 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 B8/M8-toets voor leerlingen halverwege groep 8) aan de moeilijke kant. Voor een aantal leerlingen van niveau I (of niveau A) zijn de toetsen echter aan de gemakkelijke kant. Voor leerlingen die zich minder snel of juist sneller

ontwikkelen dan de gemiddelde leerling, is het belangrijk om het niveau van de toets af te stemmen op het niveau van de leerling in plaats van op het aantal jaren onderwijs dat de leerling gevolgd heeft. Dit noemen we toetsen op maat. Zo wordt op de meest betrouwbare manier de vaardigheid van de leerling gemeten. En uiteraard is het maken van een toets op maat prettiger voor de leerlingen. Voor het toetsen op maat wordt gebruikgemaakt van de onderliggende vaardigheidsschaal. Deze schaal maakt het mogelijk om de resultaten van leerlingen die verschillende toetsen voor een bepaald leergebied maken toch met elkaar te vergelijken. Ook kan zo de ontwikkeling van individuele leerlingen in de tijd worden gevolgd. De onderliggende meettechniek voorziet er namelijk in dat iedere ruwe score – op welke toets van 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 M8-leerling maakt een toets E7) of een volgend afnamemoment (een E7-leerling maakt de toets B8/M8).

*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 toetsmap bevat hiervoor een afnamekaart met afname-instructies en een of twee voorbeeldopgaven. De leerkracht/IB'er leest vervolgens één voor één de 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 opgeschreven woorden nog een keer na. Daarna haalt de leerkracht/IB'er de antwoordbladen op.
De leerkracht/IB'er kan ervoor kiezen om de toets individueel af te nemen bij leerlingen met concentratieproblemen, leerlingen die langzamer dan gemiddeld werken of bij leerlingen die afwezig waren bij de klassikale afname. Belangrijk is dat de leerkracht of IB'er zich ook bij een individuele afname aan de afname-instructies houdt.
In de toetsmap 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 nakijkinstructies. Indien gewenst kan de leerkracht/IB'er in het Computerprogramma LOVS 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/IB'er kan ook alleen het aantal goede antwoorden invoeren in het Computerprogramma LOVS. Ook dan wordt de toetsscore 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 of IB'er de fouten van leerlingen verder wil analyseren, kan hij gebruikmaken 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 betreffende spellingcategorie(ën) ingedeeld. Er wordt aangegeven of het om de beoogde categoriefout gaat of om een andere categorie en om welke categorie het dan gaat. Het kan voorkomen dat een gemaakte fout niet ingedeeld kan worden bij een categorie. In dat geval wordt de fout aangegeven als 'andere fout'.

Op schoolniveau kan een IB'er en/of directeur met de computer een dwarsdoorsnede en trendanalyses opvragen. Met behulp van deze overzichten kan de effectiviteit van het gegeven onderwijs op groeps- en schoolniveau geanalyseerd worden.

In de handleiding bij de toetsmap worden in hoofdstuk 4 de interpretatie- en analysemogelijkheden op leerling- en groepsniveau behandeld. In hoofdstuk 5 van de handleiding komt de interpretatie op schoolniveau aan bod. De handleiding gaat in op de inhoudelijke interpretatie van de rapportages. In de handleiding bij het Computerprogramma LOVS staan de aanwijzingen over de wijze waarop de rapportages zijn op te vragen en welke keuzemogelijkheden de school hierbij heeft.
In de toetsmaterialen zijn twee niveau-indelingen opgenomen, waarmee de leerkracht de scores van een leerling kan vergelijken met die van een grote groep leerlingen (zie ook hoofdstuk 2.3). De leerkracht kan een keuze maken uit een indeling in de niveaus:
– I tot en met V;
– A tot en met E.

Daarnaast heeft de leerkracht de mogelijkheid om functioneringsniveaus op te vragen. De functioneringsniveaus geven aan met welke gemiddelde leerling de vaardigheidsscore van de getoetste leerling vergelijkbaar is. Een functioneringsniveau M8 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling overeenkomt met de score van de gemiddeld scorende leerling medio groep 8. 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 8 (hoofdstuk 6). Daarin staat een uitgebreide beschrijving van de methode-analyses en een overzicht per toets van alle getoetste woorden.

### 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. 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 en de werkwoordspelling).

In hoofdstuk 2 van deze wetenschappelijke verantwoording 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 (fonologisch, orthografisch, lexicaal-morfologisch, morfologisch op syntactische basis en logografisch) 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 methodeanalyse hebben we de indeling van spellingcategorieën vervolgens verder verfijnd tot een in- en verdeling van spellingcategorieën over alle toetsen Spelling 3.0.

Hieronder wordt de totstandkoming en onderbouwing van deze categorieënoverzichten 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 de spellingleergangen van verschillende taalmethoden met elkaar vergeleken.

In tabel 3.1 staan de methoden die in de analyse zijn meegenomen.

*Tabel 3.1 Onderzochte methoden*

| **Methode** | **Uitgever** | **Jaar van uitgave** |
|---|---|---|
| Spelling in beeld | Uitgeverij Zwijsen B.V., Tilburg | 2006, 2013 |
| Spelling op maat | Noordhoff Uitgevers, Houten | 2006, 2013 |
| Staal | Uitgeverij Malmberg, 's-Hertogenbosch | 2013 |
| Taal actief | Uitgeverij Malmberg, 's-Hertogenbosch | 2003, 2012 |
| (Taaljournaal | Uitgeverij Malmberg, 's-Hertogenbosch | 2005) |
| (Taalleesland | ThiemeMeulenhoff, Utrecht | 2004-2007) |
| Taalverhaal | ThiemeMeulenhoff, Utrecht | 2002-2003, 2013 |
| Zin in spelling | Uitgeverij Zwijsen B.V., Tilburg | 2006 |

Bij een aantal methoden staat het jaar van uitgave van twee edities. In 2010 hebben we de eerdere edities bekeken, in 2013-2015 hebben we in een aanvullende methode-analyse de nieuwste edities bekeken. De definitieve gegevens van de methodeanalyse voor de toetsen Spelling niet-werkwoorden zijn gebaseerd op beide edities.
Bij de analyse van werkwoordspelling hebben we de twee methodes die in tabel 3.1 tussen haakjes staan niet meegenomen. Deze waren inmiddels uit het assortiment van de uitgevers gehaald. Voor de vaststelling van het categorieënoverzicht voor werkwoordspelling hebben we ons beperkt tot de meest recente uitgaven van de methoden.

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 methoden het meest gebruikt worden in het basisonderwijs ten tijde van de peiling. In dit rapport was 'Staal' nog niet opgenomen. Deze methode is namelijk in 2013 voor het eerst verschenen. In ons aanvullend onderzoek van 2013 hebben we deze methode wel meegenomen. 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. Voor de toetsen Spelling niet-werkwoorden geldt, dat een categorie wordt opgenomen als de betreffende spelling-moeilijkheid in tenminste vijf van de acht methoden aan bod was geweest. Voor de toetsen Spelling werkwoorden geldt, dat een categorie wordt opgenomen, wanneer deze in tenminste vier van de zes onderzochte methoden aan bod is geweest. 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 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 8*
Alle spellingcategorieën zijn uiteindelijk bij elkaar gezet in een categorieënoverzicht, één voor niet-werkwoorden en één voor werkwoorden. Hierin is aangegeven op welk(e) moment(en) een bepaalde categorie aan bod komt in de toetsen Spelling. Als categorieën in meerdere toetsen aan bod komen, is dit duidelijk zichtbaar.
In bijlage 1 van deze verantwoording is een uitgebreid categorieënoverzicht opgenomen van de categorieën voor Spelling niet-werkwoorden die aan bod komen in de toetsen van groep 3 tot en met 8.
In bijlage 2 van deze verantwoording is een categorieënoverzicht opgenomen van de categorieën voor werkwoordspelling in groep 7 en 8. In de categorieënoverzichten in bijlage 1 en 2 is te zien dat per toets slechts een deel van het totaal aantal spellingcategorieën aan de orde komt. In de handleiding bij het toetspakket staan de precieze aantallen per categorie en per taak in paragraaf 6.3.

**Opbouw overzicht spellingcategorieën niet-werkwoorden**
Het categorieënoverzicht in bijlage 1 is opgebouwd uit een aantal kolommen (verticale indeling). Horizontaal gezien is het overzicht ingedeeld in verschillende 'blokken' die door lijntjes van elkaar gescheiden zijn. We bespreken het overzicht eerst per kolom en daarna per 'blok'.

*Per kolom*
In de eerste kolom staat het nummer van de spellingcategorie. De categorieën zijn van boven naar beneden oplopend genummerd. De nummering van categorieën sluit dankzij de methodeanalyse globaal aan op de volgorde waarin deze spellingmoeilijkheden in het onderwijs worden aangeboden. De nummering wordt op twee plekken onderbroken: na categorie 21 en na categorie 32. Na categorie 21 volgt categorie 7+, daarna 8+ en zo verder tot categorie 32. Na categorie 32 volgen de categorieën 20++ en 21++, daarna 24+ en zo verder tot categorie 47. Hierna (in het stukje 'Per blok') wordt hier nader op ingegaan.

In de tweede kolom staat de omschrijving van de spellingcategorie. Sommige categorieën komen meer dan een keer in het overzicht voor, bijvoorbeeld categorie 14 (woorden met (-)ei(-) of (-)ij(-)). Deze categorie komt dan terug als categorie 14+.

In de derde kolom staat het aantal lettergrepen van de woorden in de categorie. De eerste categorieën zijn gevuld met éénlettergrepige woorden. Vanaf categorie 8 komen ook woorden van twee lettergrepen voor en vanaf categorie 7+ kunnen ook woorden met drie of meer lettergrepen aan bod komen. Het is niet zo dat vanaf een bepaalde categorie alleen nog langere woorden gevraagd worden. Een éénlettergrepig woord als 'hardst' hoort beslist niet thuis in de lagere categorieën.

In de vierde kolom staan enkele voorbeeldwoorden. Deze woorden zijn bedoeld ter illustratie van de spellingkwestie die in de categorie aan de orde komt. Ook geven de voorbeeldwoorden een indicatie van de moeilijkheid van de woorden die in de categorie vallen.

In de volgende kolommen staat in welke toetsen Spelling niet-werkwoorden de betreffende categorie aan bod komt.

*Per blok*
In de laatste kolommen, 'In LVS-toets Spelling 3.0', is te zien dat de toetsen trapsgewijs verspringen ten opzichte van elkaar. Op twee plaatsen in het overzicht wordt een relatief grote sprong gemaakt: tussen de toetsen E4 en M5 en tussen de toetsen E6 en M7. Op deze overgangen staat in het overzicht een dubbele horizontale streep. Door de strepen wordt het overzicht in drie blokken verdeeld: een blok voor de toetsen van groep 3 en 4, een blok voor de toetsen van groep 5 en 6 en een blok voor de toetsen van groep 7 en 8. Door de indeling in blokken ontstaat de mogelijkheid om een bepaalde categorie in een groot aantal toetsen te gebruiken en daarbij toch recht te doen aan de opklimmende moeilijkheidsgraad van woorden binnen die categorie. Hiervoor maken we gebruik van plustekens. Bij het begin van het tweede blok begint de nummering als het ware opnieuw: na categorie 21 komt categorie 7+. Een dergelijk verschijnsel doet zich opnieuw voor bij de overgang naar het derde blok: na categorie 32 komt categorie 20++.

Het plusteken geeft aan dat:
– de betreffende categorie al in een vorig blok aan de orde is geweest;
– het gaat om dezelfde spellingkwestie (de omschrijving is in sommige gevallen wel iets aangepast, zie bijvoorbeeld categorie 15 en 15+);
– het gaat om woorden die langer en/of moeilijker zijn qua spelling dan de woorden in het vorige blok.

Categorienummers zonder plus komen in een bepaald blok voor het eerst voor. De categorieën in het eerste blok hebben dus nooit een plus. Als een categorienummer gevolgd wordt door een plus, weet je dat de betreffende categorie thuishoort in ofwel het tweede ofwel het derde blok. Categorienummers zonder plus komen in alle drie de blokken voor. Immers ook in groep 7 en 8 komen er nog spellingcategorieën voor de eerste keer aan de orde.

*Twee varianten van het categorieënoverzicht*
Het werken met plussen bij categorienummers heeft als voordeel dat binnen een categorie een gradatie kan worden aangebracht naar woordlengte en/of moeilijkheid. De categorieën komen dan twee keer of zelfs drie keer in het overzicht voor.
Een nadeel van deze werkwijze is dat het overzicht in de lengte flink wordt uitgerekt. Daardoor lijkt het ook of de overgang van E4 naar M5 zeer abrupt verloopt, terwijl juist deze overgang erg vloeiend is. Er komen immers in de toets M5 slechts twee nieuwe categorieën bij in vergelijking met toets E4, te weten de categorieën 22 en 23. Alle andere categorieën in de toets M5 zijn ook al in groep 4 aan de orde geweest (en sommige zelfs ook al in groep 3).
Omwille van de overzichtelijkheid hebben we er daarom voor gekozen om in de toetspakketten voor leerkrachten geen onderscheid te maken tussen categorieën met en zonder plus. Daardoor is het categorieënoverzicht in de toetsmap beknopter. In het toetspakket voor groep 8 zijn drie aparte

categorieënoverzichten opgenomen voor Spelling niet-werkwoorden, als bijlage bij de handleiding: één voor de toetsen van groep 3 en 4, één voor de toetsen van groep 5 en 6 en één voor de toetsen van groep 7 en 8. We verwijzen in de toetsmap naar het categorieënoverzicht in de onderhavige wetenschappelijke verantwoording voor meer informatie over de opklimmende woordlengte en moeilijkheid van de woorden binnen een categorie.

**Opbouw overzicht spellingcategorieën werkwoorden**
In bijlage 2 is het categorieënoverzicht voor de toetsen Spelling werkwoorden 3.0 van groep 7 en 8 opgenomen. In dit overzicht is aangegeven op welk(e) moment(en) een spellingcategorie aan bod komt in de toetsen Spelling werkwoorden.
De categorieën voor de werkwoordspelling zijn geordend in vier groepen, afhankelijk van de grammaticale categorie van het woord waarin de fout gemaakt wordt: onvoltooid tegenwoordige tijd, onvoltooid verleden tijd, voltooide deelwoorden en bijvoeglijk gebruikte deelwoorden. Voor bijzondere spellingmoeilijkheden binnen deze categorieën zijn er subcategorieën zoals ‘wel of geen -t achter een stam op -d’ binnen categorie 1 ‘onvoltooid tegenwoordige tijd’.
De spellingcategorieën die zijn opgenomen in de toetsen Spelling werkwoorden 3.0 komen grotendeels overeen met de uitwerking die Van der Beek en Paus (2011) geven van het Referentiekader Taal en Rekenen (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008a). Om beter aan te sluiten bij de praktijk van het spellingonderwijs en bij de moeilijkheden die leerlingen daadwerkelijk ondervinden bij het spellen, hebben we in enkele gevallen een categorie toegevoegd, opgesplitst of weggelaten. De in het Referentiekader genoemde inhoud is daarmee optimaal gedekt.
Voor de verleden tijd is een categorie toegevoegd voor sterke werkwoorden die in de 2e en 3e persoon enkelvoud eindigen op *-d*, om fouten als *werdt* en *hieldt* op te sporen. De categorie zwakke werkwoorden is in de verleden tijd opgesplitst in drie categorieën: zwakke werkwoorden die een stam hebben op *-v* of *-z*, zwakke werkwoorden die een stam hebben op *-d* of *-t*, en de overige zwakke werkwoorden zonder speciale moeilijkheid.
Bij het voltooid deelwoord is er een aparte categorie voor zwakke werkwoorden die speciale moeilijkheden opleveren, met een stam op *-v* of *-z* en *-d* of *-t*, naast een categorie voor de overige werkwoorden. Ook is de categorie ‘(on)voltooid deelwoord bijvoeglijk gebruikt’ toegevoegd. Bijvoeglijk gebruikte voltooid deelwoorden volgen weliswaar ‘gewoon’ de spellingregels voor voltooide deelwoorden, terwijl onvoltooide deelwoorden de regels volgen voor onveranderlijke woorden, ze hebben echter voldoende bijzondere eigenschappen om een aparte categorie te rechtvaardigen. Bovendien wordt aan bijvoeglijk gebruikte deelwoorden apart aandacht besteed in veel spellingmethoden.
Ten slotte zijn er ook categorieën niet opgenomen: ‘infinitief’ en ‘tegenwoordige tijd meervoud’. De lastigste vorm van infinitieven was oorspronkelijk wel opgenomen in onze proeftoetsingen, namelijk de infinitieven waarvoor een gelijkklinkende verleden tijdsvorm bestaat, zoals wieden (dat net zo klinkt als wiedden) en raden (dat net zo klinkt als raadden). Leerlingen maakten daarbij echter zeer weinig fouten, en bovendien zagen we dat de leerlingen die over het algemeen het best konden spellen, juist bij deze opgaven wel fouten gingen maken. Het opnemen van de categorie ‘infinitief’ in de toetsen levert dus geen informatie op over welke leerlingen beter en welke leerlingen minder goed spellen. Daarom is deze categorie in de uiteindelijke toetsen niet opgenomen. Om dezelfde reden is er geen categorie ‘tegenwoordige tijd meervoud’ opgenomen. De uitgang *-en* in de onvoltooid tegenwoordige en verleden tijd levert namelijk in de praktijk weinig problemen op voor de meeste leerlingen. Spellingmoeilijkheden in de tegenwoordige tijd doen zich vooral voor bij de enkelvoudsvormen, specifiek bij vormen eindigend op *-d* en *-t*.

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

*Itemconstructie*
Alle opgaven die in de toetsen Spelling 3.0 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 8, is telkens nagegaan of het woord voorkwam in Digiwak. Digiwak is een digitale woordenlijst. Deze bevat

alle woorden die leerlingen van het primair onderwijs in een bepaalde groep zouden moeten beheersen. De woorden zijn ingedeeld naar thema en naar groep (gebaseerd op Kuiken & Droge (2010)).
De constructiegroepleden maakten dicteezinnen bij de woorden. Zij kregen de instructie de dicteezinnen zo kort en concreet mogelijk te maken. Een toetsdeskundige heeft waar nodig zinnen aangepast om hieraan te voldoen. Op deze manier komen we tegemoet 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 van de toetsen Spelling niet-werkwoorden 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 zinnen als: ‘Wie is er aan de beurt? Schrijf op: beurt’.
De leerling heeft dan de neiging om antwoord te geven op de vraag. Of: ‘Ga eens rechtop zitten! Schrijf op: rechtop’. Ook in dat geval kan de leerling zich aangesproken voelen. Zinnen in de ik-vorm zijn bij voorkeur niet opgenomen, omdat met name leerlingen met een stoornis in het autistische spectrum deze zinnen op zichzelf kunnen betrekken. Bij de toetsen Spelling werkwoorden konden we de ik-vorm, gebiedende wijs en vraagzinnen niet vermijden, omdat bepaalde spellingregels niet anders getoetst kunnen worden.

*Proeftoetsing en kalibratie-analyses*
De opgaven zijn eerst in proefafnames voorgelegd aan leerlingen in groep 8. Het doel van dergelijke proefafnames is het verkrijgen van informatie over de moeilijkheid van elke opgave. Ook 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 de opzet van de proeftoetsingen is ervoor gekozen om een taak met nieuwe opgaven te koppelen aan de reguliere LVS-afname Spelling. Zo viel het voor de leerlingen niet op dat ze aan een proeftoets deelnamen en konden de leerkrachten gewoon de ontwikkeling van hun leerlingen blijven volgen.

Bij de proeftoetsing voor Spelling niet-werkwoorden zijn in januari 2015 een aantal opgaven voorgelegd aan leerlingen van groep 8. Daarbij zijn op het afnamemoment midden groep 8 (januari) 153 nieuwe opgaven geproeftoetst. Elke deelnemende school maakte op het medio-moment één taak met 35 nieuwe opgaven, naast de LVS-toets Spelling van de tweede generatie. De nieuwe opgaven waren verdeeld over 7 verschillende boekjes. Hierbij was sprake van een onvolledig, maar verbonden design. Vrijwel alle nieuwe opgaven werden door minimaal 200 leerlingen gemaakt.

De proeftoetsing voor de toets Spelling werkwoorden voor groep 8 is uitgevoerd in januari 2016. 127 nieuwe opgaven werden geproeftoetst. Elke deelnemende school maakte op het medio-moment één taak met 30 nieuwe opgaven, naast de LVS-toets Spelling werkwoorden van de tweede generatie. De nieuwe opgaven waren verdeeld over 7 verschillende boekjes. Hier was sprake van een onvolledig, maar verbonden design. Ook nu werden vrijwel alle nieuwe opgaven door minimaal 200 leerlingen gemaakt.

Na de afnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programmapakket One Parameter Logistic Model (OPLM; Verhelst, 1993; Verhelst en Glas, 1995). Zie voor een algemene technische beschrijving van dit model paragraaf 2.4.2.
Bij de analyses is de kwaliteit van de afzonderlijke items en 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. Men vond de opgaven ook geschikt voor het SBO. Enkele opgaven waar praktijkdeskundigen opmerkingen bij hadden, hebben we waar mogelijk aangepast (een andere dicteezin geconstrueerd) of verwijderd.

*Aanvullende proeftoetsing Spelling niet-werkwoorden*
Bij analyse van de proeftoetsresultaten van Spelling niet-werkwoorden bleek dat voor een aantal spellingcategorieën onvoldoende geschikte items overbleven, doordat veel opgaven te gemakkelijk bleken (p-waarde hoger dan .90). Het gaat om de volgende spellingcategorieën:
26 (woorden waarin /s/ geschreven wordt als c)
27 (woorden waarin /k/ geschreven wordt als c)
28 (woorden beginnend met 's of eindigend op 's)
35 (woorden met een trema)
37 (samenstelling met tussen -e(n)-)
41 (woorden waarin /t/ geschreven wordt als th)
43 (woorden waarin /ks/ geschreven wordt als x)
46 (woorden op -iaal, -ieel, -ueel, -eaal)
47 (stoffelijke bijvoeglijke naamwoorden)

Al deze spellingcategorieën worden (vrij) laat in het onderwijs aangeboden, waardoor toetsing in groep 8 wenselijk is. Deze spellingmoeilijkheden hebben echter gemeen, dat ze vrij beperkte groepen woorden betreffen. Een gevolg hiervan is, dat soms bijna de gehele groep van woorden die deze spellingmoeilijkheid bevatten, al uitputtend geïnstrueerd en geoefend is op school, waardoor leerlingen deze specifieke woorden erg goed kennen. Bijvoorbeeld bij de categorie van stoffelijke bijvoeglijke naamwoorden op -en, zoals gouden, zinken en papieren. Nagenoeg alle leerlingen spellen deze woorden goed. Bij de spelling-moeilijkheden als 'woorden met th' en 'woorden met x' lopen we tegen een variant van dit probleem aan: het Nederlands heeft daar weliswaar een ruim aantal woorden in beschikbaar, maar veel van deze woorden zijn niet passend in de leefwereld van kinderen (bv. hypotheek), of zijn onbekend of ongeschikt voor leerlingen van deze leeftijd (bv. extract). Het resultaat is dat ook voor deze categorieën het aantal voor toetsing geschikte woorden klein is en al vaak sterk ingeoefend op school.
Om het tekort in de betreffende categorieën aan te vullen zijn in het voorjaar van 2016 33 opgaven in bovenstaande categorieën voorgelegd aan 98 leerlingen. Daardoor waren er voor de selectie voor het normeringsonderzoek extra opgaven beschikbaar.

*Normeringsonderzoek M8*
Op basis van de psychometrische analyses en de evaluaties van de proeftoetsingen hebben we de opgaven geselecteerd voor de normeringsonderzoeken van januari 2017. 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 8. 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.
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 ($R_{ir}$ vanaf .20) kwamen in principe in aanmerking voor opname in de normeringsonderzoeken Spelling.

De opgaven die na de proefafnames geselecteerd waren, 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. Deze taken bevatten 30 opgaven voor Spelling niet-werkwoorden en 25 opgaven voor Spelling werkwoorden. 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 die op elkaar lijken zijn zo veel mogelijk verspreid binnen een taak. 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 werden in de normeringsonderzoeken reserve-opgaven meegenomen. Op die manier kunnen eventueel slecht functionerende opgaven nog vervangen worden door andere in de uitgavetaken.
Zie voor uitgebreide informatie over de opzet en de resultaten van de normeringsonderzoeken M8 hoofdstuk 4.

*Samenstelling definitieve toetsen*
Na de normeringsonderzoeken is van alle opgaven opnieuw de p-waarde en de $R_{ir}$ 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 ($R_{ir}$ vanaf .20) in aanmerking voor opname in de definitieve toetsen Spelling.
Verder is bij Spelling niet-werkwoorden opnieuw gekeken naar de indeling in spellingcategorieën. In de meeste dicteewoorden in de toets Spelling niet-werkwoorden van groep 8 komt 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. Een woord als 'lancering' is bijvoorbeeld ingedeeld bij categorie 26 (woorden waarin /s/ wordt geschreven als c) en niet bij categorie 20++ (woorden met open lettergreep). Om de definitieve categorie van een woord vast te stellen, hebben we bij het normeringsonderzoek geïnventariseerd hoe vaak een bepaalde spellingfout daadwerkelijk gemaakt werd door leerlingen. Hiervoor gebruikten we frequentielijsten: lijsten waarin alle verschillende gegeven antwoorden in het normeringsonderzoek, voorzien van frequentie van vóórkomen, opgesomd zijn. Voor de woorden die we hebben opgenomen in de definitieve toets, 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 een fout maakt in een andere spelling-categorie dan de beoogde categorie. Met een foutenanalyse (handmatig of via het Computerprogramma LOVS) heeft de leerkracht de mogelijkheid een diepgaandere analyse uit te (laten) voeren.
Voor de woorden in de toets Spelling werkwoorden geldt dat ze maar bij één categorie ingedeeld kunnen worden. Deze categorieën zijn dan ook niet meer aangepast na de onderzoeken.

Bij de samenstelling van de definitieve toets Spelling niet-werkwoorden groep 8 bleek het aantal items verlaagd te kunnen worden tot 25 items per taak. Er was dan nog steeds een goede vertegenwoordiging van alle spellingcategorieën. Items die in psychometrisch opzicht minder goed functioneerden, konden daarom uit de toetsen worden verwijderd. Dit betreft items met een te lage $R_{ir}$, of een te hoge of juist een te lage p-waarde.
Bij de aanpassing is rekening gehouden met een mogelijk volgorde-effect: de volgorde van de opgaven in de definitieve toets wijkt zo min mogelijk af van die van het normeringsonderzoek. Wanneer volgorde-effecten van de wijzigingen te verwachten waren, bijvoorbeeld doordat de taak nu zou beginnen of eindigen met een moeilijk item, of wanneer twee op elkaar lijkende items dicht bij elkaar zouden terechtkomen, zijn de betreffende items op een andere plaats gezet.

Bij de samenstelling van de definitieve toets Spelling werkwoorden is het aantal items gehandhaafd op 25 per taak. In de definitieve toets vielen opgaven af vanwege een te lage $R_{ir}$, een te hoge of juist een te lage p-waarde. Soms vielen opgaven af die psychometrisch gezien goed functioneerden, maar die tot een spellingcategorie behoorden die al voldoende vertegenwoordigd was in de toets. Daarentegen werden soms opgaven gehandhaafd die eigenlijk wat te moeilijk of te makkelijk waren, maar waarvoor in de betreffende spellingcategorie geen beter functionerende alternatieven voorhanden waren. Bij elke individuele opgave vond dus een afweging plaats op zowel inhoudelijke als psychometrische gronden.

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.

Het bleek bij zowel de toets Spelling niet-werkwoorden als de toets Spelling werkwoorden mogelijk om aan alle drie criteria te voldoen.
De uiteindelijke verdeling van aantallen opgaven per categorie per afnamemoment is een zo goed mogelijk compromis tussen eisen van inhoudelijke aard en psychometrische kwaliteit. Hierbij willen we graag benadrukken dat alle items, ongeacht de spellingcategorie waarop ze betrekking hebben, goed passen op de onderliggende vaardigheidsschaal van spelling niet-werkwoorden of spelling werkwoorden.

De toetsen bevatten opgaven van uiteenlopende moeilijkheidsgraad. De toetsen zijn hierdoor geschikt om verschillen tussen leerlingen in beeld te brengen. Een goede illustratie hiervan en van de samenstelling van de toetsen zijn de figuren in bijlage 3: p50- en p80-kanspunten van de opgaven in de toetsen voor groep 8 in relatie tot de gemiddelde vaardigheidsscore voor de afnamemomenten. In deze figuren is zichtbaar dat de toetsen opgaven bevatten van uiteenlopende moeilijkheidsgraad. In de figuren is de verdeling van de opgaven over de taken van de toetsen visueel weergegeven. De balkjes in de figuren geven het p50- (onderkant van het balkje) en p80-kanspunt (bovenkant van het balkje) van elke opgave aan. Het p50-punt geeft de vaardigheidsscore aan waarbij er sprake is van een kans van 50% om een opgave goed te beantwoorden.

Bij de toets Spelling niet-werkwoorden B8/M8 zijn er makkelijke opgaven (die liggen onder de stippellijn van M8), opgaven van gemiddelde moeilijkheid (doorkruisen de lijn van M8) en moeilijke opgaven (liggen boven de lijn van M8). De meeste opgaven hebben een gemiddelde moeilijkheid. Ook zijn er naar verhouding veel opgaven relatief gemakkelijk; dat is prettig voor leerlingen die minder sterk zijn in spelling. Maar ook de betere leerlingen kunnen laten zien wat ze in huis hebben door de moeilijke opgaven.
Een vergelijkbaar beeld is te zien bij de toets Spelling werkwoorden B8/M8: ook deze toetsen hebben een spreiding van makkelijke opgaven, opgaven van gemiddelde moeilijkheid en moeilijke opgaven. Ook hier zien we dat de meeste opgaven een gemiddelde moeilijkheid hebben.

Bij de analyses is de kwaliteit van de afzonderlijke items en de totale verzameling voor een afnamemoment in kaart gebracht. Itemparameters zijn geschat. Bij de analyses van de antwoorden van de leerlingen op de opgaven is nagegaan of de verschillende onderdelen (van niet-werkwoorden cq. werkwoorden) 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. Voorafgaand aan deze uitgebreide beschrijving geven we hier een globaal overzicht van de toetsen voor groep 8. In tabel 3.2a worden de beschrijvende gegevens van de toetsen Spelling niet-werkwoorden van groep 8 gepresenteerd, zowel op de ruwe scoreschaal als op de vaardigheidsschaal. De gegevens van de genormeerde toets Spelling niet-werkwoorden zijn gebaseerd op 1914 leerlingen voor M8.

In tabel 3.2b worden de beschrijvende gegevens van de toets Spelling werkwoorden van groep 8 gepresenteerd, ook zowel op de ruwe scoreschaal als op de vaardigheidsschaal. De gegevens van de genormeerde toets Spelling werkwoorden zijn gebaseerd op 1560 leerlingen voor M8.

*Tabel 3.2a Beschrijvende gegevens toets Spelling niet-werkwoorden groep 8 op de ruwe scoreschaal en op de vaardigheidsschaal*

| | Gemiddelde | Standaarddeviatie | Kurtosis | Scheefheid |
|---|---|---|---|---|
| M8 nww Ruwe score | 33,8 | 9,5 | -0,52 | -0,42 |
| M8 nww Vaardigheid | 366,8 | 27,1 | 0,06 | 0,24 |

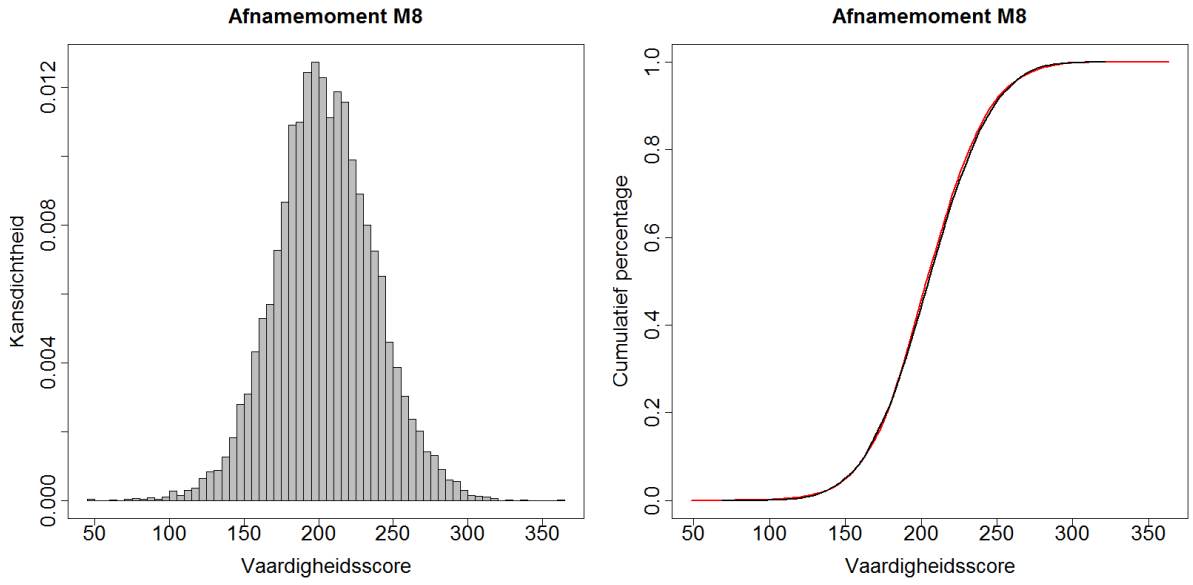
*Tabel 3.2b* *Beschrijvende gegevens toets Spelling werkwoorden groep 8 op de ruwe scoreschaal en op de vaardigheidsschaal*

| | Gemiddelde | Standaarddeviatie | Kurtosis | Scheefheid |
|---|---|---|---|---|
| M8 ww Ruwe score | 34,9 | 8,3 | -0,70 | -0,21 |
| M8 ww Vaardigheid | 142,7 | 23,4 | 0,28 | 0,24 |

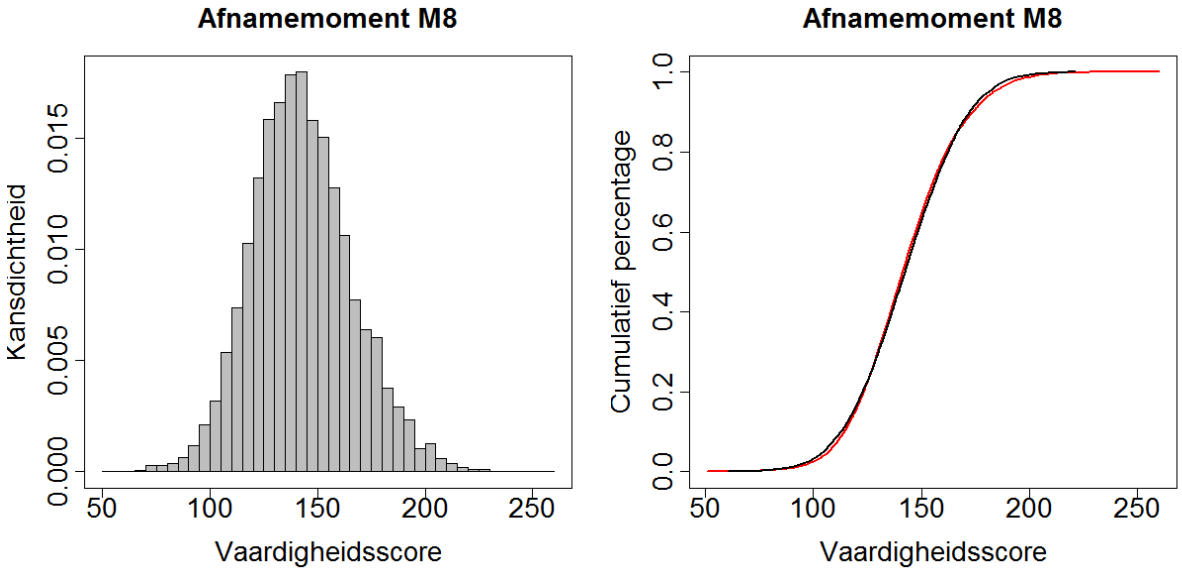
De ruwe scores zijn licht scheef verdeeld. Dit is niet verwonderlijk. De toetsen worden samengesteld rond een verwachte (gewenste) p-waarde van .70.
Voor de vaardigheidsverdelingen, de scores die gebruikt worden om leerlingen te vergelijken en te volgen, wordt in principe uitgegaan van een normale verdeling, zoals te zien is in het linkerpaneel van de figuren 3.1 en 3.2. De rechterfiguur geeft de cumulatieve verdeling weer alsmede de cumulatieve verdeling behorend bij de normaalverdeling. Deze verdelingen ontlopen elkaar zeer weinig.

Het verschil tussen de vaardigheidsscores zoals weergegeven in de figuren 3.1 en 3.2 en die in tabel 3.2a en 3.2b is te verklaren doordat de scores in de figuren worden weergegeven op de kalibratieschaal voor groep 8 terwijl de gemiddelden en standaarddeviaties in tabel 3.2a en 3.2b op de (getransformeerde) overkoepelende vaardigheidsschaal worden weergegeven (zie voor een toelichting hoofstuk 4).

*Figuur 3.1* *Weergave van behaalde vaardigheidsscores en cumulatieve verdeling voor afnamemoment M8 niet*-werkwoorden



*Figuur 3.2* *Weergave van behaalde vaardigheidsscores en cumulatieve verdeling voor afnamemoment M8 werkwoorden*



# 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 M8 moeten vergelijkbaar zijn met die van andere afnamemomenten, bijvoorbeeld M7 en E7. Met andere woorden, het dataverzamelingsdesign dient verbonden te zijn. Hiertoe dient een longitudinale opzet gebruikt te worden.

*De verbondenheid van het design*
Het idee van een gemeenschappelijke schaal impliceert strikt genomen dat men iemands vaardigheid zou kunnen schatten aan de hand van een willekeurig samengestelde toets. Het spreekt echter vanzelf dat het een zinloze onderneming is een toets die geconstrueerd is voor groep 8 voor te leggen aan leerlingen van groep 4, omdat zo'n toets ongetwijfeld opgaven zal bevatten die een beroep doen op kennis van leerstof die in groep 4 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 E7 bijvoorbeeld betekent dit dat de leerlingen in het kalibratie- en normeringsonderzoek items krijgen voorgelegd die specifiek voor E7 zijn geconstrueerd, en (een minderheid aan) items die geconstrueerd zijn voor M7 en M8. Voor M8 betekent dit dat de leerlingen in het kalibratie- en normeringsonderzoek items voorgelegd krijgen die specifiek voor M8 zijn geconstrueerd, en (een minderheid aan) items die geconstrueerd zijn voor E7.

Het macro-design van Spelling niet-werkwoorden is weergegeven in figuur 4.1[5]. Het macro-design van Spelling werkwoorden is een stuk kleiner en is weergegeven in figuur 4.3.

[5] B8 maakt geen onderdeel uit van het macro-design. De opzet van het normeringsonderzoek B8 wordt apart beschreven in een nog te verschijnen addendum.

*Figuur 4.1 Macro-design LVS-toetsen Spelling 3.0*

| | | M3 | | E3 | | M4 | | E4 | | M5 | | E5 | | M6 | | E6 | | M7 | | E7 | | M8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| jan 2013 | M3 en 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 |

De items die voor de overlap of verankering zorgen, duiden we in het macro-design aan met ank, gevolgd door 2 cijfers. Zo duidt ank78 de groep items aan die enerzijds bestaat uit items geconstrueerd voor E7 en anderzijds uit items geconstrueerd voor M8. Die items zijn dus zowel eind groep 7 als medio groep 8 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 M3, E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8.

*Longitudinale opzet van Spelling niet-werkwoorden*
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 5 op het eindmoment van schooljaar x en neemt eveneens deel aan de opvolgende momenten M6 (schooljaar x+1) en E6 (schooljaar x+1). School B start op moment M6 (schooljaar Y) en neemt eveneens deel aan de opvolgende momenten E6 (schooljaar y) en M7 (schooljaar y+1). Op deze manier wordt tegemoetgekomen aan 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 wordt de kalibratie per leerjaar uitgevoerd op een beperkt deel van de gemeenschappelijke schaal. De kalibratie vindt plaats op basis van de verzamelde data voor dat leerjaar op de afnamemomenten, aangevuld met de gegevens van het voorgaande en het opvolgende afnamemoment. Omdat M8 het sluitstuk vormt, is in het geval van leerjaar 8 geen aparte schaal gemaakt voor groep 8. Dat betekent dat de kalibratie met afnamemoment M8 plaatsvindt op basis van afnamemomenten E6, M7, E7 en M8. Deze opzet sluit aan bij de inhoudelijke kenmerken van de aangeboden opgaven. 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 is telkens op een gedeelte van het eerder vermelde macro-design uit figuur 4.1 gefocust. In het geval van groep 8 betreft het dus het gedeelte dat in figuur 4.2 hieronder is weergegeven. Zoals te zien is in de wetenschappelijke verantwoording van Spelling 3.0 groep 7 (Tomesen, Wouda, Krämer & Horsels, 2018) komt de kalibratie van de toets Spelling niet-werkwoorden van groep 8 exact overeen met de kalibratie van groep 7.

*Figuur 4.2 Gedeelte macro-design waarop kalibratie Spelling niet-werkwoorden M8 is gebaseerd*

| | | E6 | | M7 | | E7 | | M8 |
|---|---|---|---|---|---|---|---|---|
| Juni 2015 | ank66 | E6 | ank67 | | | | | |
| Januari 2016 | | | ank67 | M7 | ank77 | | | |
| Juni 2016 | | | | | ank77 | E7 | ank78 | |
| Januari 2017 | | | | | | | ank78 | M8 |

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 beschrijven we in de volgende paragrafen.

Om de prestaties van leerlingen en groepen te kunnen blijven volgen, zullen deze op een overkoepelende schaal worden geplaatst door gebruik te maken van een transformatie. Deze transformatie wordt afgeleid uit de overlappende populaties op de kalibraties. De overlappende jaargroepen op opvolgende schalen bestaan uit dezelfde leerlingen in beide kalibraties en hebben per definitie dezelfde vaardigheidsverdeling. Om deze reden kan uit de vaardigheidsverdelingen van die jaargroepen de transformatie berekend worden.

*Macro-design van Spelling werkwoorden*
Gezien het kleinere aantal meetmomenten van Spelling werkwoorden, kan worden volstaan met een zeer ingeperkt macro-design. Dit design wordt afgebeeld in figuur 4.3.

*Figuur 4.3 Macro-design waarop kalibratie Spelling werkwoorden M8 is gebaseerd*

| | M7 | | E7 | | M8 |
|---|---|---|---|---|---|
| Januari 2017 | M7 | ank77 | | | |
| Juni 2016 | | ank77 | E7 | ank78 | |
| Januari 2017 | | | | ank78 | M8 |

Voor Spelling werkwoorden zijn er alleen toetsen en afnamemomenten op M7, E7 en M8[6], er zijn geen eerdere of latere toetsen, respectievelijk afnamemomenten. Transformatie naar een overkoepelende lange-termijn-schaal, zoals bij Spelling niet-werkwoorden (en andere LVS-toetsen) is daarom niet nodig. Zoals te zien is in de wetenschappelijke verantwoording van Spelling 3.0 groep 7 (Tomesen, Wouda, Krämer & Horsels, 2018) komt de kalibratie van de toets Spelling werkwoorden van groep 8 exact overeen met de kalibratie van groep 7.
Afgezien van bovenstaande verschil is de opzet van het normeringsonderzoek van Spelling werkwoorden gelijk aan dat van Spelling niet-werkwoorden.

## 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. In principe gelden alle beschrijvingen van de kalibratie voor zowel Spelling niet-werkwoorden als Spelling werkwoorden. Mocht de procedure afwijken voor Spelling werkwoorden, dan zal dit expliciet vermeld worden.

### 4.2.1 De opzet van de kalibratie

Prestaties van leerlingen blijken al snel na publicatie van een toets te verschuiven, omdat bij het onderzoek dat ten grondslag ligt aan de normering sprake is van low stakes afnamesituaties (Keuning et al., 2015). Bij de ontwikkeling van 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

[6] Er is ook een afnamemoment B8, maar die valt buiten dit macro-design. Het normeringsonderzoek B8 wordt apart beschreven in een nog te verschijnen addendum.

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 de reguliere afname van de 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 bij het normeringsonderzoek M8: 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[7] van de tweede generatie LVS-toetsen (zie Keuning et al., 2015).

In figuur 4.4 en 4.5 worden de *embedded field* designs weergegeven voor respectievelijk M8 niet-werkwoorden en M8 werkwoorden. In het design kunnen we zien dat er voor niet-werkwoorden zes toetsversies zijn afgenomen voor M8. Voor werkwoorden zijn er vijf toetsversies afgenomen voor M8. 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[8].

*Figuur 4.4 Design Spelling 3.0 M8 niet-werkwoorden*

| Toets-versie | M8 LVS 2e gen. | M8 deel 1 | M8 deel 2 | E7 anker 1 | E7 anker 2 | Reservetaak | Passieve taak Taalverzorging | Leerlingen | Scholen |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ■ | ■ | ■ | | | | | 191 | 9 |
| 2 | ■ | | | ■ | | ■ | | 195 | 9 |
| 3 | ■ | | ■ | | | | ■ | 77 | 5 |
| 4 | ■ | ■ | | | | ■ | | 154 | 8 |
| 5 | ■ | | | | ■ | | ■ | 172 | 8 |
| 6 | ■ | | ■ | ■ | | | | 167 | 8 |

[7] Cito dataretour is een exporttool die basisscholen in staat stelt om jaarlijks op vrijwillige basis hun LVS-resultaten naar Cito te sturen voor (interne) onderzoeksdoeleinden. Het opsturen van resultaten vindt geautomatiseerd plaats via het Computer-programma LOVS. Verreweg de meeste basisscholen geven gehoor aan de oproep die Cito jaarlijks doet. Het grote voordeel van Cito dataretour is dat er longitudinale toetsgegevens beschikbaar komen van honderdduizenden leerlingen op verschillende toetsen.

[8] Met uitzondering van de boekjes 3 en 5 van M8 niet-werkwoorden en de boekjes 3 en 4 van M8 werkwoorden. De leerlingen die deze boekjes hadden, maakten naast de taak van de LVS-toets Spelling uit de tweede generatie één taak van Spelling 3.0 én een taak passieve spelling uit de LVS-toets Taalverzorging.

*Figuur 4.5 Design Spelling 3.0 M8 werkwoorden*

| Toets-versie | M8 LVS 2e gen. | M8 deel 1 | M8 deel 2 | E7 anker | Reservetaak | Passieve taak Taalverzorging | Leerlingen | Scholen |
|---|---|---|---|---|---|---|---|---|
| 1 | ■ | ■ | ■ | | | | 160 | 9 |
| 2 | ■ | | ■ | | ■ | | 187 | 9 |
| 3 | ■ | | | | ■ | ■ | 131 | 6 |
| 4 | ■ | | | ■ | | ■ | 212 | 9 |
| 5 | ■ | ■ | | ■ | | | 182 | 9 |

De taken M8 deel 1 en M8 deel 2 vormen tezamen de beoogde uitgave Spelling 3.0 M8. Dit geldt voor zowel Spelling niet-werkwoorden als Spelling werkwoorden.
In de normeringsonderzoeken is een taak met ‘reserve-opgaven’ meegenomen. Deze opgaven zouden in de uiteindelijke uitgave alleen worden gebruikt indien er onverwachte problemen naar voren kwamen met betrekking tot de beoogde opgaven voor de uitgaven M8.
Sommige leerlingen maakten op afnamemoment M8 in plaats van twee taken van de beoogde uitgave Spelling één taak van de beoogde uitgave én een taak ‘passieve spelling’ uit de toets Taalverzorging. Dat hebben we gedaan om de samenhang te kunnen documenteren tussen actieve spelling en passieve spelling. Zie meer daarover in hoofdstuk 6.

Zoals te zien in de designs vormden de toetsen M8 uit de tweede generatie van het LVS een stevig anker tussen de toetsboekjes en werd er tussen de beoogde toetsen voor Spelling 3.0 geankerd in de normeringsonderzoeken. Er is dus 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 8 niet-werkwoorden en werkwoorden. Ook kunnen we door deze opzet de normering van de nieuw uit te geven toetsen vergelijken met de normering van de tweede generatie en kan de continuïteit tussen de tweede en derde generatie in beeld worden gebracht (zie hoofdstuk 6 over validiteit).

*Aantallen items en leerlingen normeringsonderzoek Spelling groep 8 niet-werkwoorden*
In het normeringsonderzoek M8 in januari 2017 zijn 198 verschillende items voorgelegd aan 993 leerlingen van groep 8. Hiervan hebben 956 leerlingen de volledige taken gemaakt. Dit aantal leerlingen is verdeeld over 6 boekjes, zoals aangegeven in de voorlaatste kolom van figuur 4.4. Elk boekje bestond uit 78-90 opgaven, verdeeld over 3 taken. De 30 opgaven uit de bestaande taak Spelling van de tweede generatie werden door alle 956 leerlingen uit het normeringsonderzoek gemaakt. De overige opgaven kwamen steeds in 2 boekjes voor en werden gemiddeld door 330 leerlingen gemaakt. Op grond van de gegevens uit de kalibratie van de normeringsonderzoeken is de definitieve selectie van items gemaakt voor de uitgave van de toets Spelling niet-werkwoorden 3.0 M8.

*Aantallen items en leerlingen normeringsonderzoek Spelling groep 8 werkwoorden*
In het normeringsonderzoek M8 in januari 2017 zijn 144 verschillende items voorgelegd aan 890 leerlingen van groep 8. Hiervan hebben 872 leerlingen de volledige taken gemaakt. Dit aantal leerlingen is verdeeld over 5 boekjes, zoals aangegeven in de voorlaatste kolom van figuur 4.5. Elk boekje bestond uit 70-75 opgaven, verdeeld over 3 taken. De 25 opgaven uit de bestaande taak Spelling van de tweede generatie werden door alle 872 leerlingen uit het normeringsonderzoek gemaakt. De overige opgaven kwamen steeds in 2 boekjes voor en werden gemiddeld door 352 leerlingen gemaakt. Op grond van de gegevens uit de kalibratie van de normeringsonderzoeken is de definitieve selectie van items gemaakt voor de uitgave van de toets Spelling werkwoorden 3.0 M8.

### 4.2.2 De stappen in de kalibratie

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

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

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

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

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

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

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

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

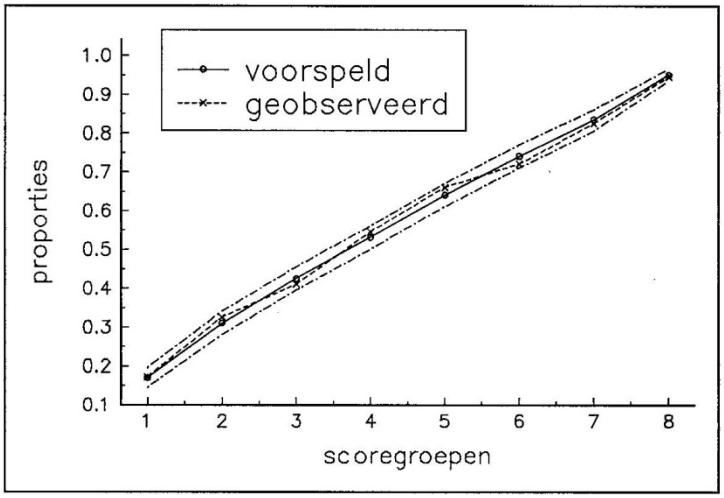
4. Een volgende controle betreft de overschrijdingskansen van de S-toetsen en een grafische modelcontrole door middel van het programma WOPPLOT (grafische inspectie van de ICC's).
5. Vervolgens vindt een globale modelcontrole plaats in de vorm van een R1c-toets en de verdeling van de overschrijdingskansen 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 4.6 (zie Staphorsius, 1994, blz. 239). Figuur 4.6 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., 1995).

*Figuur 4.6 Grafische voorstelling van een Si –toets*

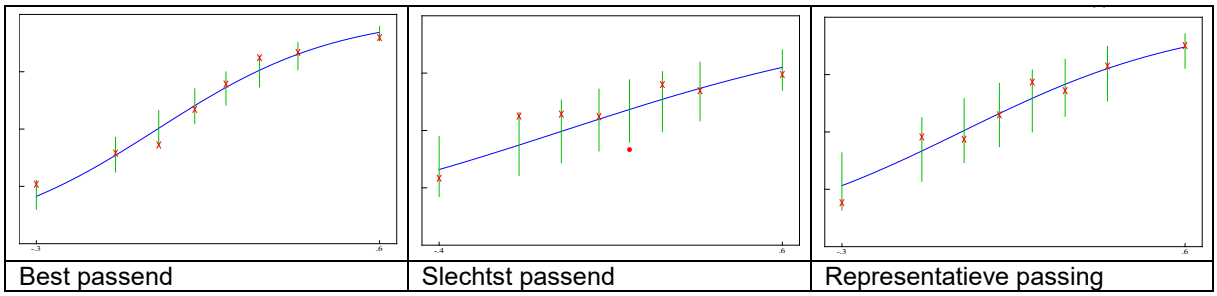



Het is ondoenlijk om voor alle opgaven dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons steeds per toetsversie tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.7 illustreren dat voor de opgaven van Spelling groep 8 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 4.6 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 4.7a Voorbeelden van S-toetsen voor de toets Spelling 3.0 B8/M8 niet-werkwoorden met de best passende, de slechtst passende en een qua passing representatieve opgave*

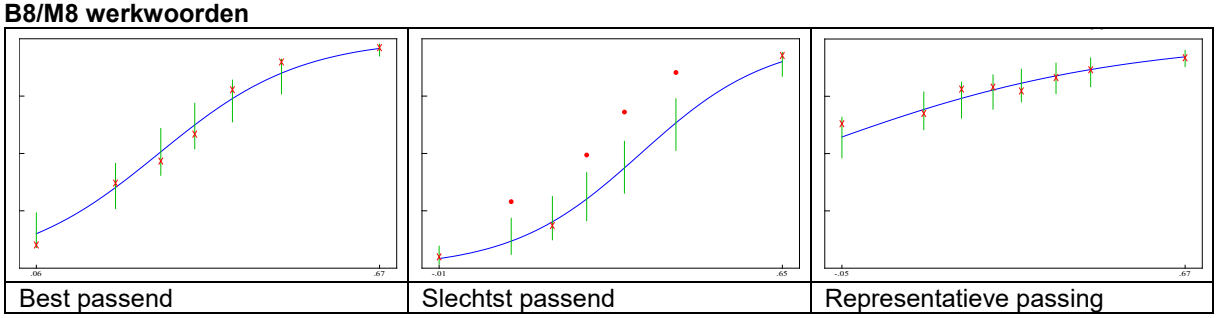
**B8/M8 niet-werkwoorden**



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

*Figuur 4.7b Voorbeelden van S-toetsen voor de toets Spelling 3.0 B8/M8 werkwoorden met de best passende, de slechtst passende en een qua passing representatieve opgave*

**B8/M8 werkwoorden**



| 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 8. 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. Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensionaal construct representeren.

*Tabel 4.1a Verdeling van overschrijdingskansen bij S-toetsen voor toets Spelling 3.0 B8/M8 niet-werkwoorden*

| | 0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B8/M8 | 0 | 0 | 1 | 8 | 6 | 3 | 3 | 2 | 5 | 6 | 7 | 9 |

*Tabel 4.1b Verdeling van overschrijdingskansen bij S-toetsen voor toets Spelling 3.0 B8/M8 werkwoorden*

| | 0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B8/M8 | 0 | 3 | 1 | 2 | 3 | 5 | 3 | 6 | 1 | 7 | 8 | 11 |

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 groter dan ongeveer anderhalf maal het aantal vrijheidsgraden (df) is. Voor beide toetsen geldt dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootheid is significant (bij α = 0,05) voor B8/M8 niet-werkwoorden. Aan dit laatste moet bij steekproeven met een dergelijke omvang niet te veel waarde worden gehecht.

*Tabel 4.2a R1c-waarden voor toets Spelling 3.0 B8/M8 niet-werkwoorden*

| Toetsversie | R1c | df | p |
|---|---|---|---|
| B8/M8 | 684,2 | 515 | <0,005 |

*Tabel 4.2b R1c-waarden voor toets Spelling 3.0 B8/M8 werkwoorden*

| Toetsversie | R1c | df | p |
|---|---|---|---|
| B8/M8 | 394,5 | 377 | 0,26 |

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 0,20. Waarden tussen 0,30 en 0,40 kunnen nog als voldoende worden beschouwd.

In tabel 4.3a en 4.3b zijn gemiddelde en range van deze waarden voor alle toetsitems weergegeven.
De gemiddelde waarde van de constante is goed te noemen voor de toets Spelling 3.0 groep 8 niet-werkwoorden. Voor alle items op twee na geldt dat de waarden kleiner zijn dan 0,20. De gemiddelde waarde voor de toets Spelling 3.0 werkwoorden is ook goed. Een achttal waarden zijn iets boven de 0,20, maar deze blijven allemaal ruimschoots onder 0,30.

*Tabel 4.3a Nauwkeurigheid van de itemparameterschattingen voor toets Spelling 3.0 B8/M8 niet-werkwoorden (constante 'c')*

| Toetsversie | Constante 'c' | |
|---|---|---|
| | range | gemiddelde |
| B8/M8 | 0,066-0,224 | 0,124 |

*Tabel 4.3b Nauwkeurigheid van de itemparameterschattingen voor toets Spelling 3.0 B8/M8 werkwoorden (constante 'c')*

| Toetsversie | Constante 'c' | |
|---|---|---|
| | range | gemiddelde |
| B8/M8 | 0,058-0,240 | 0,131 |

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen Spelling 3.0 B8/M8 niet-werkwoorden en werkwoorden 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 de toetsen 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. (2015). Allereerst besteden we aandacht aan de opzet van het normeringsonderzoek, de gehanteerde procedures en de aantallen leerlingen per afnamemoment (paragraaf 4.3.1). Vervolgens komt in paragraaf 4.3.2 de representativiteit van de normsteekproeven aan de orde. De paragraaf wordt afgerond met een presentatie van de resultaten van de normering (i.e. de kenmerken van de 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 werd 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. Dit betekent dat representatieve steekproeven zoals deze hier zijn ingezet voor de normering van de toetsen Spelling 3.0 in groep 8 normeringsresultaten zullen opleveren die niet of nauwelijks zullen afwijken van wat men voor de totale populatie van scholen zou mogen verwachten. In eerdere normeringsonderzoeken voor groep 3 en 4 is dat voor meerdere jaargangen nagegaan. Niet alleen de gemiddelde score op de Cito Eindtoets Basisonderwijs, maar ook specifiek de score voor Spelling bleek voor de scholen in deze normerings-onderzoeken niet af te wijken van het populatiegemiddelde voor de Eindtoets.

Scholen is gevraagd op drie opeenvolgende momenten deel te nemen aan de normeringsonderzoeken Spelling niet-werkwoorden en Spelling werkwoorden. Van de deelnemende scholen op afnamemoment E7 zijn scholen benaderd als herhalingsscholen voor het normeringsonderzoek op afnamemoment M8.
De normeringsgroep van afnamemoment M8 bestond dus deels uit herhalingsscholen.
Voor het totaal aantal scholen in Nederland (6681 scholen) is een indeling gemaakt naar LOVS-strata[9] (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 5%. Op basis van deze schattingen werd het aantal aan te schrijven scholen bij de werving vastgesteld.

Voor de normeringsonderzoeken M8 niet-werkwoorden en M8 werkwoorden zijn in eerste instantie 79 herhalingsscholen (scholen die ook meededen aan normeringsonderzoek E7) aangeschreven. Omdat na de eerste aanschrijvingsronde 42% (niet-werkwoorden) respectievelijk 39% (werkwoorden) van de herhalingsscholen uit normeringsonderzoek E7 bereid bleek deel te nemen aan het normeringsonderzoek M8, zijn in een tweede aanschrijvingsronde 2245 scholen aangeschreven. Uiteindelijk resulteerde dit in een deelnamebereidheid van zo'n 1% van de overige aangeschreven scholen. In totaal meldden zich 57 scholen aan voor het normeringsonderzoek niet-werkwoorden, waarvan uiteindelijk 52 scholen daadwerkelijk gegevens aanleverden. Voor het normeringsonderzoek werkwoorden meldden zich 51 scholen in totaal, waarvan uiteindelijk 46 scholen daadwerkelijk gegevens aanleverden. Van een aantal scholen zijn vervolgens de data niet meegenomen in de kalibratie en normering, omdat bleek dat deze scholen de toetsen niet hadden afgenomen volgens de afnamecondities.

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 voor wie het onderwijstype onbekend is;
– leerlingen van scholen die het LVS selectief inzetten. In de hogere leerjaren blijken sommige scholen het LVS namelijk alleen in te zetten bij zwakkere leerlingen (zie Keuning, 2011);
– leerlingen die op hetzelfde afnamemoment meerdere toetsen van dezelfde vaardigheid maken. Alleen de gegevens van de toets die bij het afnamemoment hoorde, werden behouden. Daarnaast werden de scholen verwijderd die ook aan de *embedded field* normeringsonderzoeken deelnamen.

[9] De term stratum wordt hier gedefinieerd zoals gebruikelijk in periodieke peilingsonderzoeken, namelijk als een indicatie van de aard van de schoolpopulatie in termen van het percentage leerlingen met een afwijkend leerlingewicht per school.

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* normerings-onderzoek 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 wordt rekening gehouden met de variabelen *regio*, *urbanisatiegraad*, *schooltype* en *sekse*. De verschillende variabelen zijn als volgt gedefinieerd:

– ***Regio***. Bij de definitie van de variabele *regio* is uitgegaan van de CBS-indeling naar landsdeel. Dit betekent dat er vier regio’s onderscheiden zijn. Regio *noord* omvat de provincies Groningen, Friesland en Drenthe; regio *oost* de provincies Overijssel, Gelderland en Flevoland; regio *west* de provincies Utrecht, Noord-Holland, Zuid-Holland en Zeeland en regio *zuid* de provincies Noord-Brabant en Limburg.

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

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

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

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

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

Tot nu toe is bij de selectie van data uitsluitend rekening gehouden met de schoolvariabelen *regio*, *urbanisatiegraad* en *schooltype.* De leerlingvariabele *sekse* is nog niet in beschouwing genomen. Dat gebeurt in de tweede stap. Als blijkt dat de normeringssteekproef die is samengesteld in de eerste stap niet representatief is met betrekking tot de variabele *sekse*, dan wordt een tweede steekproeftrekking uitgevoerd.

Eerst wordt op basis van de landelijke gegevens van CBS en de geobserveerde aantallen in de normeringssteekproef de kans $w_q$ bepaald dat een leerling met sekse $q$ in een representatieve normeringssteekproef zit:

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

Zoals we kunnen zien, wordt het gewenste percentage leerlingen in categorie $q$ gedeeld door het geobserveerde percentage leerlingen in categorie $q$. Als $w_q$ voor alle leerlingen in de normeringssteek-proef 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 ($\mathrm{S}$) en Cito dataretour ($\mathrm{D}$) dus als volgt in zijn werk:

*Voorbereiding data normeringsonderzoek*

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

*Toevoegen data uit Cito dataretour*

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

*Check leerlingvariabele sekse*

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

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 8 niet-werkwoorden ingesteld op $N_T$ = 2 x 956 = 1912 en voor afnamemoment medio groep 8 werkwoorden op $N_T$ = 2 x 872 = 1744.

In de tabellen 4.4a en 4.4b is te zien in welke aantallen scholen en leerlingen het selectiealgoritme heeft geresulteerd. De conclusie is dat het zowel voor het afnamemoment medio groep 8 niet-werkwoorden als voor afnamemoment medio groep 8 werkwoorden tot de gewenste oplossing heeft geleid. De aantallen leerlingen die via het *embedded field* normeringsonderzoek en uit dataretour bij de normering zijn betrokken wijken weliswaar enigszins af van de nagestreefde 50:50 verhouding, maar dit is een gevolg van het exacte verloop van het algoritme gegeven de verdeling van scholen over de categorieën in de achtergrond-variabelen. 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 de tabellen 4.4a en 4.4b wordt weergegeven welke aantallen van de steekproef en van dataretour uiteindelijk zijn meegenomen in de normering.

*Tabel 4.4a Aantal leerlingen die meegenomen zijn in de normering Spelling niet-werkwoorden*

| Afnamemoment | Aantal leerlingen | | | Aantal scholen |
|---|---|---|---|---|
| | Steekproef | Dataretour 2e generatie | Normering | Normering |
| M8 | 945* | 969 | 1914 | 114 |

* Het aantal leerlingen kan lager zijn dan dat van de oorspronkelijke normeringssteekproef doordat sommige scholen in de toepassing van het beschreven algoritme niet zijn geselecteerd.

*Tabel 4.4b Aantal leerlingen die meegenomen zijn in de normering Spelling werkwoorden*

| Afnamemoment | Aantal leerlingen | | | Aantal scholen |
|---|---|---|---|---|
| | Steekproef | Dataretour | Normering | Normering |
| M8 | 654* | 906 | 1560 | 87 |

* Het aantal leerlingen kan lager zijn dan dat van de oorspronkelijke normeringssteekproef doordat sommige scholen in de toepassing van het beschreven algoritme niet zijn geselecteerd.

### 4.3.2 Representativiteit

Door de werkwijze die werd gevolgd tijdens de normering van M8 is representativiteit van de totale normeringssteekproeven (inclusief dataretour) voor M8 in principe gegarandeerd. Niettemin werd er een controle uitgevoerd op de representativiteit door de populatieverdelingen verkregen uit gegevens van DUO te vergelijken met de steekproefverdelingen.
In de tabellen 4.5 tot en met 4.8 worden de resultaten van de representativiteitsanalyses getoond.
De steekproef is geanalyseerd in relatie tot de variabelen regio, urbanisatiegraad, schooltype (stratum) en sekse. De aantallen in onderstaande tabellen kunnen lager zijn dan de aantallen in de normering vanwege ontbrekende achtergrondgegevens.

*Tabel 4.5a Aantal en percentage leerlingen in de populatie en de steekproef naar regio voor Spelling 3.0 groep 8 niet-werkwoorden*

| | **Populatie** | **Steekproef** | |
|---|---|---|---|
| **Regio** | **%** | **M8** | **%** |
| Noord | 9,9 | 160 | 8,4 |
| Oost | 22,2 | 427 | 22,5 |
| West | 48,2 | 944 | 49,8 |
| Zuid | 19,6 | 365 | 19,2 |

M8 $\chi^2(3, N = 1896) = 5{,}338$; $p = 0{,}149$; $\phi = 0{,}053$

*Tabel 4.5b Aantal en percentage leerlingen in de populatie en de steekproef naar regio voor Spelling 3.0 groep 8 werkwoorden*

| | **Populatie** | **Steekproef** | |
|---|---|---|---|
| **Regio** | **%** | **M8** | **%** |
| Noord | 9,9 | 119 | 7,6 |
| Oost | 22,2 | 336 | 21,5 |
| West | 48,2 | 795 | 51,0 |
| Zuid | 19,6 | 310 | 19,9 |

M8 $\chi^2(3, N = 1560) = 11{,}032$; $p = 0{,}012$; $\phi = 0{,}084$

*Tabel 4.6a Aantal en percentage leerlingen in de populatie en de steekproef naar urbanisatiegraad voor Spelling 3.0 groep 8 niet-werkwoorden*

| | **Populatie** | **Steekproef** | |
|---|---|---|---|
| **Urbanisatie** | **%** | **M8** | **%** |
| Stad | 54,3 | 1054 | 55,6 |
| Land | 45,7 | 842 | 44,4 |

M8 $\chi^2(1, N = 1896) = 1{,}300$; $p = 0{,}254$; $\phi = 0{,}026$

*Tabel 4.6b Aantal en percentage leerlingen in de populatie en de steekproef naar urbanisatiegraad voor Spelling 3.0 groep 8 werkwoorden*

| | Populatie | Steekproef | |
|---|---|---|---|
| **Urbanisatie** | **%** | **M8** | **%** |
| Stad | 54,3 | 851 | 54,6 |
| Land | 45,7 | 709 | 45,4 |

M8 $\chi^2(1, N = 1560) = 0{,}044$; $p = 0{,}834$; $\phi = 0{,}005$

*Tabel 4.7a Aantal en percentage leerlingen in de populatie en de steekproef naar stratum voor Spelling 3.0 groep 8 niet-werkwoorden*

| | Populatie | Steekproef | |
|---|---|---|---|
| **Stratum** | **%** | **M8** | **%** |
| 0 – 10% | 57,2 | 1127 | 59,4 |
| 10 – 25% | 28,4 | 540 | 28,5 |
| 25 – 40% | 7,4 | 117 | 6,2 |
| > 40% | 7,0 | 112 | 5,9 |

M8 $\chi^2(3, N = 1896) = 8{,}688$; $p = 0{,}034$; $\phi = 0{,}068$

*Tabel 4.7b Aantal en percentage leerlingen in de populatie en de steekproef naar stratum voor Spelling 3.0 groep 8 werkwoorden*

| | Populatie | Steekproef | |
|---|---|---|---|
| **Stratum** | **%** | **M8** | **%** |
| 0 – 10% | 57,2 | 942 | 60,4 |
| 10 – 25% | 28,4 | 461 | 29,6 |
| 25 – 40% | 7,4 | 73 | 4,7 |
| > 40% | 7,0 | 84 | 5,4 |

M8 $\chi^2(3, N = 1560) = 24{,}690$; $p < 0{,}0005$; $\phi = 0{,}126$

*Tabel 4.8a Aantal en percentage leerlingen in de populatie en de steekproef naar sekse voor Spelling 3.0 groep 8 niet-werkwoorden*

| | Populatie | Steekproef | |
|---|---|---|---|
| **Geslacht** | **%** | **M7** | **%** |
| Jongen | 50,5 | 898 | 48,6 |
| Meisje | 49,5 | 951 | 51,4 |

M8 $\chi^2(1, N = 1849) = 2{,}665$; $p = 0{,}103$; $\phi = 0{,}038$

*Tabel 4.8b Aantal en percentage leerlingen in de populatie en de steekproef naar sekse voor Spelling 3.0 groep 8 werkwoorden*

| | Populatie | Steekproef | |
|---|---|---|---|
| **Geslacht** | **%** | **M7** | **%** |
| Jongen | 50,5 | 768 | 51,3 |
| Meisje | 49,5 | 729 | 48,7 |

M8 $\chi^2(1, N = 1497) = 0{,}420$; $p = 0{,}517$; $\phi = 0{,}017$

De $\chi^2$-waarden zijn over het algemeen laag, maar in een aantal gevallen significant. Bij grotere steekproeven zegt significantie echter niet zoveel. Het is beter om de effectgrootte $\phi$ als uitgangspunt te nemen.

*Formule 4.1 Berekening van de effectgrootte $\phi$*

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

We zien dat de effectgroottes over het algemeen ver onder of rond de 0,10 liggen en daarmee klein tot zeer klein zijn (cf. Cohen, 1988) voor afnamemoment M8. De conclusie is dat de normeringssteekproeven een goede afspiegeling vormen van de populatie. Om deze reden is statistische weging van de resultaten van de normeringssteekproef niet nodig.

### 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., 2015). 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 afnamemoment M8 te zien. De *plausible values* vormen een normale verdeling.

Op basis van deze scoreverdeling worden de percentielen berekend die horen bij de vaardigheids-indelingen A tot en met E en I tot en met V zoals beschreven in hoofdstuk 2. Daarbij wordt uitgegaan van de empirische cumulatieve verdelingsfunctie. Voor Spelling niet-werkwoorden geldt dat deze percentielen volgens een transformatie op de overkoepelende schaal over leerjaren geplaatst worden. Tabellen 4.9a en 4.9b geven de normgegevens voor respectievelijk de toetsen Spelling 3.0 groep 8 niet-werkwoorden en werkwoorden.

*Tabel 4.9a Normtabel op leerlingniveau voor Spelling 3.0 groep 8 niet-werkwoorden*

| Afname-moment | M | SD | Kur-tosis | Skew-ness | P10 | P20 | P25 | P40 | P50 | P60 | P75 | P80 | P90 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M8 | 366,7 | 27,1 | 0,41 | 0,00 | 332,8 | 344,9 | 349,0 | 359,5 | 366,0 | 373,3 | 384,7 | 388,7 | 400,9 |

*Tabel 4.9b Normtabel op leerlingniveau voor Spelling 3.0 groep 8 werkwoorden*

| Afname-moment | M | SD | Kur-tosis | Skew-ness | P10 | P20 | P25 | P40 | P50 | P60 | P75 | P80 | P90 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M8 | 142,8 | 23,2 | 0,06 | 0,24 | 114,6 | 123,6 | 127,1 | 135,6 | 141,5 | 147,2 | 157,5 | 162,2 | 173,9 |

Naast een normering op leerlingniveau kent Cito ook een normering op schoolniveau. Om de school-verdeling te bepalen wordt het intercept-only multilevel model gebruikt met een gemiddelde per school en een variantie op school- en leerlingniveau. De schatting van het model verloopt via een bootstrap procedure. Dit betekent dat het multilevel model meerdere keren wordt geschat, steeds op basis van een andere selectie van scholen en leerlingen uit de normeringssteekproef. Bij elke replicatie wordt het aantal scholen dat geselecteerd gaat worden gelijkgesteld aan het aantal scholen dat in de normeringssteekproef zit. Vervolgens worden binnen een school leerlingen geselecteerd. Ook dit aantal leerlingen dat geselecteerd gaat worden, wordt gelijkgesteld aan het aantal leerlingen dat feitelijk op de betreffende school zit. De scholen en leerlingen worden geselecteerd met teruglegging. Als de selectie is afgerond, wordt het multilevel model geschat en de intraklassecorrelatie en het designeffect uitgerekend. Tabellen 4.10a en 4.10b laten de resultaten van de bootstrapprocedure zien. De uitkomsten zijn behoorlijk stabiel. 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 4.10a Samenvatting uitkomsten multilevel analyse Spelling 3.0 groep 8 niet-werkwoorden*

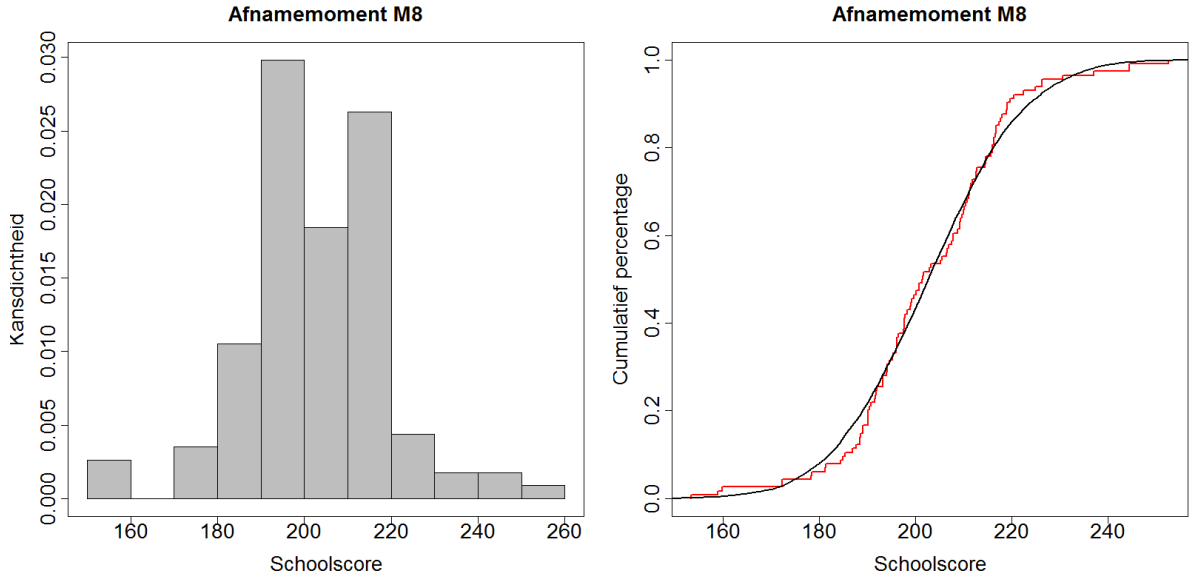
| Afname-moment | Aantal replicaties | Aantal scholen | Gemiddelde | SD School | SD Leerling | ICC |
|---|---|---|---|---|---|---|
| M8 | 20 | 114 | 203,4 | 13,6 | 33,9 | 0,14 |

*Tabel 4.10b Samenvatting uitkomsten multilevel analyse Spelling 3.0 groep 8 werkwoorden*

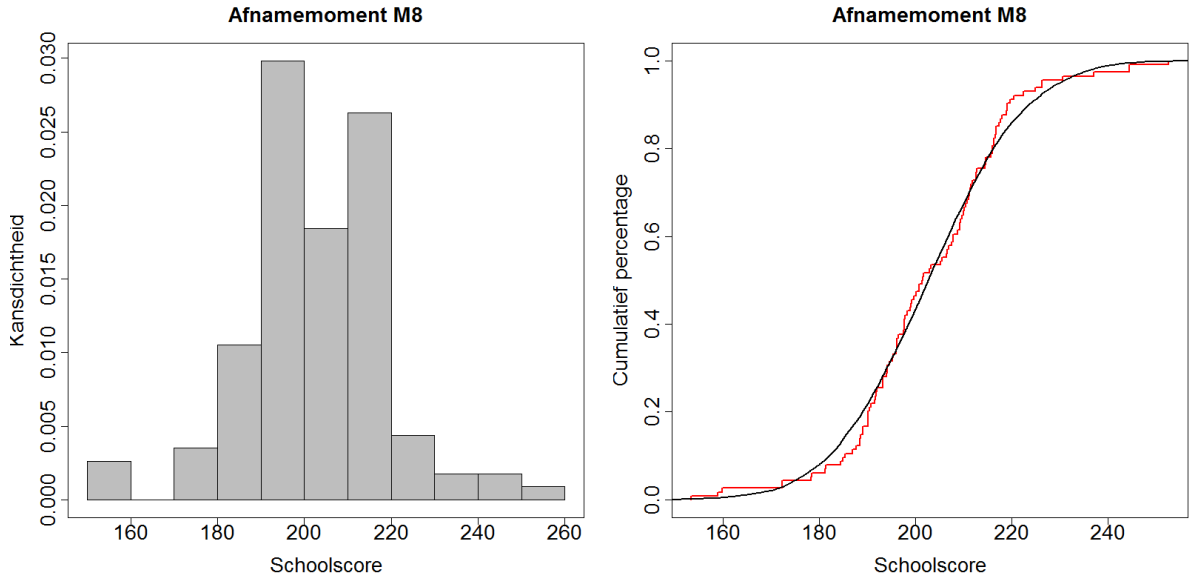
| Afname-moment | Aantal replicaties | Aantal scholen | Gemiddelde | SD School | SD Leerling | ICC |
|---|---|---|---|---|---|---|
| M8 | 20 | 87 | 142,3 | 12,2 | 26,1 | 0,18 |

Figuren 4.8 en 4.9 laten de verdelingen van schoolgemiddelden zien. Het is lastig te bepalen of de schoolgemiddelden een normale verdeling volgen met een scholenaantal van 87 en 114. Op het eerste gezicht lijkt er sprake te zijn van een normale verdeling. Op basis van het eindresultaat uit de bootstrapprocedure zijn de percentielen voor de vaardigheidsverdeling A tot en met E en I tot en met V berekend. Tabel 4.11 geeft de normgegevens op schoolniveau. De percentielen komen dichter bij elkaar te liggen dan in de leerlingverdeling. De afstanden zijn echter nog wel groot genoeg om scholen zinvol te classificeren in de verschillende niveaus.
Het verschil tussen de vaardigheidsscores zoals weergegeven in figuren 4.8 en 4.9 en tabel 4.10 en die in tabel 4.11 is te verklaren doordat de scores in de figuren 4.8 en 4.9 en tabel 4.10 worden weergegeven op de kalibratieschaal voor groep 8 terwijl de gemiddelden en standaarddeviaties in tabel 4.11 op de (getransformeerde) overkoepelende vaardigheidsschaal worden weergegeven.

*Figuur 4.8 Verdeling van de schoolgemiddelden voor Spelling 3.0 M8 niet-werkwoorden*




*Figuur 4.9 Verdeling van de schoolgemiddelden voor Spelling 3.0 M8 werkwoorden*



*Tabel 4.11a Normtabel op schoolniveau voor Spelling 3.0 groep 8 niet-werkwoorden*

| Afname-moment | M | SD | P10 | P20 | P25 | P40 | P50 | P60 | P75 | P80 | P90 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| M8 | 366,0 | 11,1 | 351,8 | 356,7 | 358,5 | 363,2 | 366,0 | 368,8 | 373,4 | 375,3 | 380,2 |

*Tabel 4.11b Normtabel op schoolniveau voor Spelling 3.0 groep 8 werkwoorden*

| Afname-moment | M | SD | P10 | P20 | P25 | P40 | P50 | P60 | P75 | P80 | P90 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| M8 | 142,3 | 12,6 | 126,7 | 132,0 | 134,1 | 139,2 | 142,3 | 145,4 | 150,5 | 152,5 | 158,0 |

### 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 8 een geldigheid aanhouden tot en met 2027. Daarnaast monitort Cito periodiek de normering. Jaarlijks wordt aan de hand van representatieve afnamedata nagegaan of er systematisch verschuivingen in het prestatieniveau plaatsvinden. Indien nodig wordt de normering aangepast.

# 5 Betrouwbaarheid en meetnauwkeurigheid

## 5.1 Betrouwbaarheid

Het is mogelijk om de betrouwbaarheid van de toetsen Spelling 3.0 voor groep 8 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 et al. (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ. Deze verwachte waarde wordt aangeduid met τ(θ). Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie wordt 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 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'.

Tabellen 5.1a en 5.1b bevatten informatie over de meeteigenschappen van de toetsen Spelling 3.0 voor groep 8. In de tweede kolom staat de maximumscore, deze is gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. De derde kolom geeft de geschatte gemiddelde score van de leerlingen op de toets. De vierde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van de toets. De vijfde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de toets 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 betrouwbaarheid van de toetsen goed te noemen.

*Tabel 5.1a Beschrijvende gegevens bij de toets Spelling 3.0 B8/M8 niet-werkwoorden*

| Afname-moment | Maximum score | Gemiddelde | Standaard-meetfout | MAcc | Test-hertest (simulatie) |
|---|---|---|---|---|---|
| M8 | 50 | 33,8 | 2,842 | 0,91 | 0,91 |

*Tabel 5.1b Beschrijvende gegevens bij de toets Spelling 3.0 B8/M8 werkwoorden*

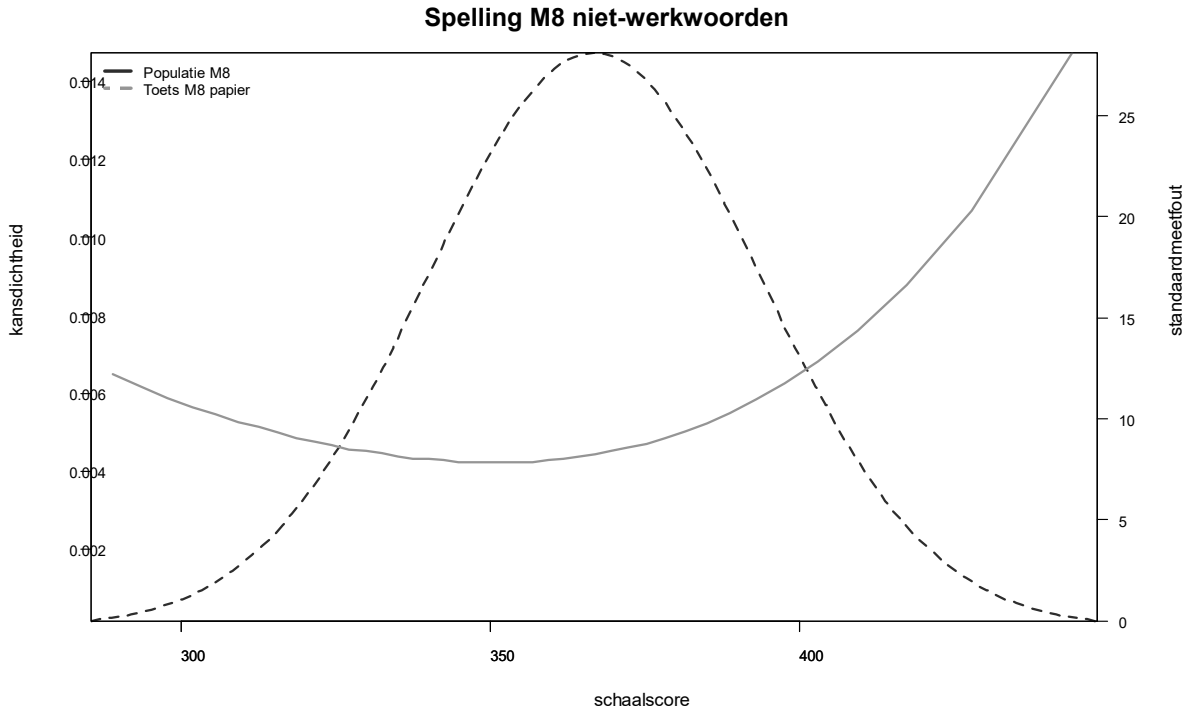
| Afname-moment | Maximum score | Gemiddelde | Standaard-meetfout | MAcc | Test-hertest (simulatie) |
|---|---|---|---|---|---|
| M8 | 50 | 34,9 | 2,914 | 0,88 | 0,88 |

Er heeft geen test-hertestonderzoek plaatsgevonden. De afnamecontext van de LVS-toetsen Spelling 3.0 leent zich daar niet goed voor. Het feit dat alle items echter OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1.000.000 leerlingen. Daarbij hebben we enerzijds de vaardigheidsverdeling van alle leerlingen, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselect uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 1.000.000 dubbele (virtuele) afnames berekend. Men kan deze simulatie beschouwen als een test-hertestonderzoek onder ideale condities. De tweede toetsafname is immers volledig onafhankelijk van de eerste toetsafname en wordt niet beïnvloed door de kennis die de leerling mogelijk verworven heeft via de eerste toetsafname. Daarnaast is er geen sprake van invloed van een test-hertestinterval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden. De resultaten zijn weergegeven in tabel 5.1a en 5.1b (zie de laatste kolom). De uitkomsten komen exact overeen met de eerder berekende coëfficiënten en deze leiden dan ook tot dezelfde conclusies met betrekking tot de betrouwbaarheid van de toetsen Spelling 3.0.

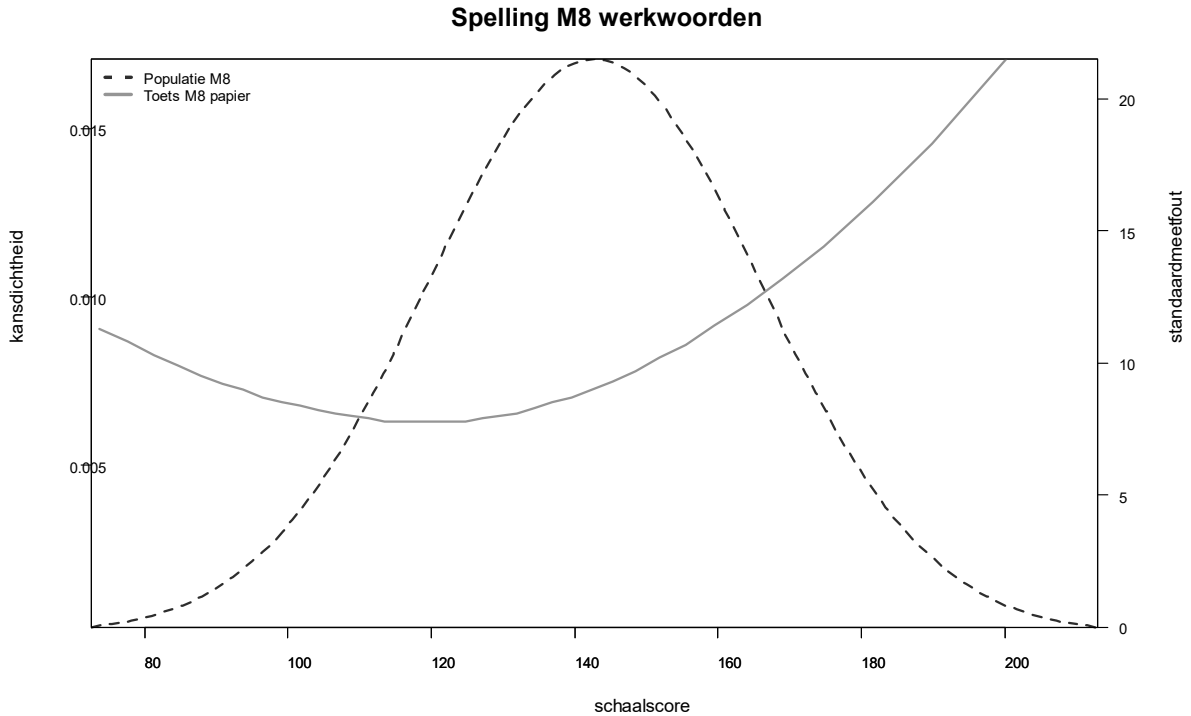
## 5.2 Nauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meet-nauwkeurigheid en geven geen zicht op de lokale meetnauwkeurigheid van de toetsen Spelling 3.0 voor groep 8 niet-werkwoorden en werkwoorden. Figuur 5.1a en 5.1b geven grafisch weer hoe het daarmee gesteld is. In deze figuren staat voor deze toetsen de grootte van de meetfout op de vaardigheidsschaal afgebeeld. Ook is de kansdichtheidsfunctie voor de normgroep op het afnamemoment opgenomen.
Deze laat zien hoe de vaardigheid van de leerlingen verdeeld is over de vaardigheidsschaal in de populatie van leerlingen waaruit een steekproef de toets gemaakt heeft. De figuur maakt duidelijk dat bij beide toetsen de meetfout kleiner is in de lagere en gemiddelde vaardigheidsregionen dan in de hogere vaardigheidsregionen.

*Figuur 5.1a Grootte van de meetfouten voor de toets B8/M8 niet-werkwoorden en de kansdichtheidsfunctie voor de M8 niet-werkwoorden-populatie*




*Figuur 5.1b Grootte van de meetfouten voor de toets B8/M8 werkwoorden en de kansdichtheidsfunctie voor de M8 werkwoorden-populatie*



**Betrouwbaarheidstabellen**

De betekenis van de (lokale) meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit betrouwbaarheidstabellen. De tabellen 5.2 en 5.3 laten voor de toets B8/M8 niet-werkwoorden en voor B8/M8 werkwoorden zien hoe vaak de werkelijke vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat tabel 5.2 zien dat 80,6 procent van de leerlingen die medio groep 8 op basis van de B8/M8-toets niet-werkwoorden in scoregroep V geclassificeerd wordt ook met hun werkelijke vaardigheidsscore in deze scoregroep ingedeeld wordt. De kans dat een V-leerling terecht als V-leerling wordt bestempeld is, met andere woorden, ongeveer 81 procent. Verder laat de linkerkant van tabel 5.2 zien dat 18,8 procent van de leerlingen in scoregroep V een vaardigheidsscore heeft die in werkelijkheid in scoregroep IV valt. De overige getallen in tabellen 5.2 en 5.3 zijn op dezelfde wijze te interpreteren.

*Tabel 5.2 Betrouwbaarheidstabel Toets B8/M8 niet-werkwoorden voor afnamemoment medio 8*

| | **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 | 80,6 | 18,8 | 0,6 | 0,0 | 0,0 | E | 78,8 | 20,9 | 0,3 | 0,0 | 0,0 |
| IV | 9,4 | 59,1 | 28 | 3,3 | 0,1 | D | 7,8 | 61,6 | 30 | 0,6 | 0,0 |
| III | 0,3 | 18,9 | 50,7 | 27,3 | 2,8 | C | 0,0 | 9,2 | 61,7 | 27,3 | 1,7 |
| II | 0,0 | 2,8 | 25,6 | 49,1 | 22,5 | B | 0,0 | 0,1 | 15,6 | 54,9 | 29,3 |
| I | 0,0 | 0,1 | 3,7 | 25,1 | 71,2 | A | 0,0 | 0,0 | 0,9 | 18 | 81,2 |

*Tabel 5.3 Betrouwbaarheidstabel Toets B8/M8 werkwoorden voor afnamemoment medio 8*

| | **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 | 79,7 | 19,3 | 0,9 | 0,0 | 0,0 | E | 75,0 | 24,0 | 0,9 | 0,0 | 0,0 |
| IV | 12,7 | 57,1 | 27,1 | 3,1 | 0,1 | D | 10,4 | 57,2 | 31,4 | 1,0 | 0,0 |
| III | 0,5 | 19,6 | 50,2 | 27,2 | 2,6 | C | 0,1 | 11,8 | 62,5 | 24,5 | 1,1 |
| II | 0,0 | 2,2 | 22,7 | 50,2 | 24,9 | B | 0,0 | 0,3 | 19,3 | 57,5 | 22,9 |
| I | 0,0 | 0,1 | 2,4 | 21,2 | 76,4 | A | 0,0 | 0,0 | 1,0 | 19,7 | 79,4 |

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. In de praktijk liggen de waarden vaak tussen 0,60 en 0,70.

De samenvattende indices voor respectievelijk niet-werkwoorden en werkwoorden zijn te vinden in de tabellen 5.4a en 5.4b. Waar de betrouwbaarheidstabellen laten zien dat de meeste leerlingen op basis van hun geschatte vaardigheidsscore geplaatst worden in de niveaugroep waar ze werkelijk thuishoren, maken tabellen 5.4a en 5.4b aannemelijk dat de uitkomsten duidelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969) of zelfs boven dit ambitieniveau uitstijgen. Gemiddeld gezien scoort, afhankelijk van de gekozen indeling in scoregroepen, 97,3 tot 99,3 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 62,1 tot 67,6 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore gemiddeld gezien in circa 65 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.

Op basis van bovenstaande gegevens concluderen we dat op basis van de toetsen Spelling 3.0 groep 8 niet-werkwoorden en werkwoorden de leerlingen op een betrouwbare manier ingedeeld kunnen worden in normgroepen. Deze indeling voldoet uitstekend gegeven het doel van de toetsen. 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 5.4a Samenvattende indices toets B8/M8 niet-werkwoorden op afnamemoment medio 8*

| | **Toets B8/M8, afnamemoment M8** | |
|---|---|---|
| | **scoregroep I t/m V** | **scoregroep A t/m E** |
| Marginal classification accuracy | 62,1 | 67,6 |
| Accuracy plus/minus 1 niveau | 97,3 | 99,3 |

*Tabel 5.4b Samenvattende indices toets B8/M8 werkwoorden op afnamemoment medio 8*

| | **Toets B8/M8, afnamemoment M8** | |
|---|---|---|
| | **scoregroep I t/m V** | **scoregroep A t/m E** |
| Marginal classification accuracy | 62,7 | 66,3 |
| Accuracy plus/minus 1 niveau | 97,7 | 99,1 |

Verdere gedetailleerde informatie over de meetnauwkeurigheid van de toets is te vinden in de handleiding van het toetspakket Spelling groep 8 (Cito, 2018). In de schaalscoretabellen van bijlage 2 in de handleiding is een kolom opgenomen waarin het score-interval vermeld is. In deze kolom staat voor iedere ruwe score op elke toets het 67-procents-betrouwbaarheidsinterval voor de bijbehorende vaardigheidsschatting.

# 6 Validiteit

De begripsvaliditeit van een toets 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 leervorderingentoetsen, zoals deze toetsen Spelling 3.0 voor groep 8, 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 toetsen 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 toetsen aangegeven en worden gegevens over de structuur van de toetsen 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 toetsen is uitgevoerd. Dit onderzoek levert data op over de convergente en divergente validiteit. Als laatste komen in paragraaf 6.6 verschillen tussen relevante groepen aan bod.

## 6.1 Inhoudsvaliditeit

De samenstelling van de toetsen is bepaald in overeenstemming met inhoudelijke 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 uitgebreide methodenanalyses, vormden de basis voor de domeinbeschrijving van de toetsen Spelling.
In hoofdstuk 3 is de domeinbeschrijving uitgewerkt in een beschrijving en verantwoording van spelling-categorieën voor spelling van niet-werkwoorden en van werkwoorden 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 onderwijs-methoden in die zin dat de getoetste stof in het merendeel van de methodes aan bod is gekomen en niet in een of enkele methodes onevenredig veel aandacht krijgt. 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 8. Aan de empirische gegevens die we hebben over de moeilijkheidsgraad is te zien dat deze inderdaad aansluit bij het ontwikkelingsniveau van leerlingen van groep 8.
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 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 4.1a en 4.1b).

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 goede modelfit geldt als belangrijkste vuistregel dat R1c bij voorkeur niet groter zou moeten zijn dan ongeveer anderhalf maal het aantal vrijheidsgraden. In tabel 4.2a en 4.2b zijn de betreffende R1c-waarden te vinden. De modelpassing van de toetsen voldoet aan de genoemde 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: één voor Spelling niet-werkwoorden en één voor Spelling werkwoorden. Dit betekent dat met elke willekeurige subset van items dezelfde onderliggende vaardigheid (spellen van niet-werkwoorden cq. spellen van werkwoorden) 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 goed 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 4.3a en 4.3b zijn gemiddelde en range van deze waarden voor alle toetsitems weergegeven. De gemiddelde waarden van de constante zijn te interpreteren als goed. Voor twee items van Spelling niet-werkwoorden en acht items van Spelling werkwoorden is c iets groter dan 0,20, wat nog altijd te kwalificeren is als voldoende. De conclusie mag luiden dat ook op basis van deze analyse de kalibratie geslaagd kan worden genoemd.

### 6.3 Itemkwaliteit

In tabel 6.1a en 6.1b zijn de ranges en de gemiddelden weergegeven voor de p-waarden en de $r_{it}$-waarden van de items van de toetsen B8/M8 van respectievelijk niet-werkwoorden en werkwoorden. Bij beide toetsen (Spelling niet-werkwoorden en Spelling werkwoorden) is te zien dat de p-waarden liggen tussen de 0,43 en 0,92. Er is gestreefd naar p-waarden van de items tussen 0,40 en 0,90. Slechts één item in de toets Spelling niet-werkwoorden heeft een iets hogere p-waarde, te weten 0,92. Er is gestreefd naar een goede spreiding van moeilijkheid over de items. De gemiddelde moeilijkheid van de toetsen Spelling 3.0 is 0,68 voor niet-werkwoorden en 0,70 voor werkwoorden.

*Tabel 6.1a Range en gemiddelde van p- en $R_{it}$-waarden voor de toets B8/M8 van Spelling 3.0 niet-werkwoorden*

| | **P-waarden** | | **$R_{it}$-waarden** | | **N items** |
|---|---|---|---|---|---|
| | **Range** | **Gemiddelde** | **Range** | **Gemiddelde** | |
| **M8** | 0,43 – 0,92 | 0,68 | 0,27 – 0,56 | 0,42 | 50 |

*Tabel 6.1b Range en gemiddelde van p- en $R_{it}$-waarden voor de toets B8/M8 van Spelling 3.0 werkwoorden*

| | P-waarden | | $R_{it}$-waarden | | N items |
|---|---|---|---|---|---|
| | Range | Gemiddelde | Range | Gemiddeld | |
| M8 | 0,43 – 0,87 | 0,70 | 0,23 – 0,57 | 0,38 | 50 |

Voor geen van de toetsen ligt de $r_{it}$-waarde onder 0,20. De gemiddelde $r_{it}$-waarde is 0,42 voor niet-werkwoorden en 0,38 voor werkwoorden. Door de Cotan wordt een $r_{it}$-waarde hoger dan 0,30 gekwalificeerd als goed. Met een gemiddelde van 0,38 en 0,42 is de itemkwaliteit van de toetsen goed te noemen. Bijlage 4 bevat een volledig overzicht van de p-waarden en de $r_{it}$-waarden van de items van de toetsen.
De eerder gememoreerde kalibratieresultaten op itemniveau en de analyses met betrekking tot de nauwkeurigheid van de itemparameterschattingen (in termen van constante 'c') sluiten aan bij de hier opgesomde positieve bevindingen.

## 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 beide toetsen was deze S-statistiek bij een enkele toetsopgave significant. De toetsen Spelling 3.0 voor groep 8 blijken dus geen of nauwelijks itembias te vertonen naar sekse; er is in de afzonderlijke items van systematische bevoordeling van jongens, dan wel meisjes geen sprake.

## 6.5 Soortgenootonderzoek – convergente en divergente validiteit

De convergente validiteit is in kaart gebracht door de samenhang te onderzoeken met de tweede generatie van de Cito LVS-toetsen Spelling als soortgenoten. Ook zijn de samenhangen onderzocht met de toetsen Spelling 3.0 niet-werkwoorden en werkwoorden voor groep 7 en met de taak passieve spelling in de Cito LVS-toetsen Taalverzorging. Bovendien hebben we gekeken naar de longitudinale vaardigheidstoename in de tijd. Daarnaast is de divergente validiteit onderzocht in termen van de samenhang met andere taalonderdelen en rekenen-wiskunde van het Cito LVS.

*Convergente validiteit: Correlatie scores Spelling 3.0 met LVS-toetsen Spelling 2e generatie*
De leerlingen van de normeringssteekproef M8 hebben zowel opgaven van de toetsen Spelling van de tweede generatie als opgaven van Spelling 3.0 gemaakt en daardoor kan de samenhang worden bepaald tussen de score op de toetsen van de tweede en de derde generatie. De latente correlatie is hoog voor het afnamemoment Spelling niet-werkwoorden M8 (r = 0,91; N = 956). De latente correlatie voor afnamemoment M8 Spelling werkwoorden (r = 0,92; N = 872) is ook hoog. Aangezien de begripsvaliditeit van de toetsen Spelling van de tweede generatie door de Cotan positief beoordeeld is, vormen deze hoge correlaties een belangrijk element in de bewijsvoering voor de validiteit van de toetsen Spelling 3.0.

*Convergente validiteit: Correlatie scores Spelling 3.0 M7 tot en met M8*
Een van de functies van de LVS-toetsen is het beschrijven en volgen van de vaardigheidsgroei in spelling van leerlingen over de jaren heen. Om een uitspraak te kunnen doen over de correlaties tussen de scores van (groepen) leerlingen op verschillende afnamemomenten is een gedeelte van de leerlingen uit de normeringssteekproef M7 (januari 2016) gevolgd. Deze leerlingen hebben in juni 2016 deelgenomen aan het normeringsonderzoek E7 en hebben in 2017 meegedaan aan het normeringsonderzoek M8 niet-

werkwoorden. Daarnaast is een gedeelte van de leerlingen uit de normeringssteekproef E7 (juni 2016) ook op moment M8 gevolgd. In tabel 6.2 worden de latente correlaties tussen de gewogen scores[10] op de verschillende afnamemomenten weergegeven voor Spelling niet-werkwoorden (correlaties onder de diagonaal; aantallen leerlingen boven de diagonaal).

*Tabel 6.2 Correlaties tussen de scores op de toetsen Spelling 3.0 M7 tot en met M8 voor Spelling niet-werkwoorden*

| | **M7** | **E7** | **M8** |
|---|---|---|---|
| **M7** | --- | 593 | 452 |
| **E7** | 0,90 | --- | 570 |
| **M8** | 0,88 | 0,90 | --- |

De correlaties voor de toets M8 niet-werkwoorden met de toetsen M7 en E7 niet-werkwoorden zijn heel hoog (0,88 resp. 0,90). Dat betekent dat over deze onderwijsperiode de spellingvaardigheid zich heeft ontwikkeld tot een zeer stabiel leerlingkenmerk. Deze resultaten wijzen er ook op dat de constructeurs erin geslaagd zijn om voor de hele periode M7-M8 toetsen te construeren die duidelijk een operationalisatie vormen van één en dezelfde vaardigheidsschaal.

De correlatie tussen de toetsen M8 werkwoorden en E7 werkwoorden was ook erg hoog (0,91; N=581). Het was niet mogelijk een correlatie te berekenen met de toets M7 werkwoorden, aangezien de normerings-steekproef van M7 ná E7 en tegelijkertijd met M8 is onderzocht en er geen sprake was van herhaalde metingen (de steekproeven bestonden uit verschillende leerlingen).

*Convergente validiteit: Correlatie scores Spelling 3.0 en onderdeel Passieve spelling uit de toets Taalverzorging*
Ook de correlatie tussen Spelling 3.0 en het onderdeel spelling (niet-werkwoorden cq. werkwoorden) uit de toets Taalverzorging van het Cito Volgsysteem primair en speciaal onderwijs is onderzocht. Dit geldt zowel voor spelling niet-werkwoorden als voor spelling werkwoorden.
In het normeringsonderzoek niet-werkwoorden medio groep 8 maakten 249 leerlingen naast twee taken spelling ook een taak spelling uit de toets Taalverzorging. De taak spelling uit de toets Taalverzorging bestond uit 20 meerkeuzeopgaven waarmee de passieve spelling wordt gemeten. Leerlingen krijgen per opgave een zin te zien met twee doelwoorden, waarbij ze moeten zoeken naar de zin waarbij beide doelwoorden goed gespeld zijn. De correlatie tussen de scores op de taken niet-werkwoorden was hoog (r = 0,82; N = 249). Deze bevinding sluit aan bij het uitgangspunt dat het bij deze toetsen om gelijkaardige vaardigheden gaat, namelijk spelling.

In het normeringsonderzoek werkwoorden medio groep 8 maakten 343 leerlingen naast twee taken spelling ook een taak spelling werkwoorden uit de toets Taalverzorging. De taak spelling werkwoorden uit de toets Taalverzorging bestond – net als die van niet-werkwoorden – uit 20 meerkeuzeopgaven waarmee de passieve spelling wordt gemeten. Ook hier krijgen leerlingen per opgave een zin te zien met twee doelwoorden, waarbij ze moeten zoeken naar de zin waarbij beide doelwoorden goed gespeld zijn.
Ook tussen de scores op de taken werkwoorden was de correlatie hoog (r = 0,84; N = 343). Hiermee kunnen we concluderen dat het bij beide vormen van werkwoordspelling om gelijkaardige vaardigheden gaat, namelijk werkwoordspelling. Deze resultaten bieden hiermee ondersteuning voor de validiteit van de toetsen Spelling.

[10] Gewogen scores zijn in het geval van OPLM gelijk aan aantal goed maal de a-parameter. Dit geldt als een *sufficient statistic* voor de vaardigheidsschatting.

*Convergente validiteit: Longitudinale vaardigheidstoename*

Een laatste, meer indirecte aanwijzing voor de convergente validiteit vinden we in de longitudinale vaardigheidstoename in de tijd. Wanneer de spellingtoetsen voor elkaar opvolgende afnamemomenten immers valide operationalisaties zijn van de latente trek die zij beogen te meten, dan is te verwachten dat de gemiddelde vaardigheid over leerjaren en afnamemomenten toeneemt. Treffen we deze vaardigheidsgroei inderdaad empirisch aan in een consistent patroon, dan vormt dit evidentie voor de validiteit van alle toetsen, ook die van groep 8. In tabel 6.3 staat een overzicht van de vaardigheidsverdeling per normeringsmoment van Spelling niet-werkwoorden. We nemen voor alle normeringsmomenten (M3 t/m M8) de schatting van het gemiddelde, de standaarddeviatie en het aantal leerlingen in de normeringssteekproef op. Deze gegevens komen uit deze en uit onze eerdere wetenschappelijke verantwoordingen van de LVS-toetsen Spelling 3.0.

*Tabel 6.3a Overzicht van de vaardigheidsverdelingen per normeringsmoment Spelling niet-werkwoorden*

| Normerings-moment | Aantal leerlingen | Aantal scholen | Gemiddelde vaardigheid | Standaard-deviatie |
|---|---|---|---|---|
| M3 | 2964 | 133 | 150,0 | 50,0 |
| E3 | 2996 | 132 | 200,4 | 40,4 |
| M4 | 6717 | 264 | 239,0 | 40,5 |
| E4 | 6094 | 220 | 263,7 | 37,7 |
| M5 | 3602 | 160 | 296,1 | 38,6 |
| E5 | 3285 | 158 | 311,9 | 38,3 |
| M6 | 3053 | 138 | 318,8 | 32,7 |
| E6 | 2836 | 132 | 333,7 | 31,6 |
| M7 | 2342 | 123 | 351,3 | 32,3 |
| E7 | 2579 | 137 | 356,7 | 28,5 |
| M8 | 1914 | 114 | 366,8 | 27,1 |

De gemiddelde vaardigheid van de leerlingen blijkt van afname tot afname toe te nemen. Dat impliceert dat de methodenanalyse die ten grondslag lag aan de samenstelling van de toetsen een correct beeld heeft gegeven van de wijze waarop het spellingonderwijs in de regel vormgegeven is. De toename in vaardigheid wordt overigens in de regel kleiner naarmate leerlingen in een hogere groep komen en vlakt sterk af in groep 7 en 8. Dat is ook wel te verwachten; vaak gaat het bij spelling in de hogere groepen om voortbouwen op de kennis die er al is. In een aantal gevallen wordt een categorie herhaald en is er daarbij alleen een toename in de moeilijkheid van de woorden. Dat is een verschil ten opzichte van de lagere groepen, waar alles nog nieuw is en van het begin af aan geleerd moet worden. De vaardigheid neemt dus nog steeds wel toe tussen twee afnamemomenten, alleen niet meer zo veel. Dat zou ook kunnen wijzen op een plafondeffect: steeds meer leerlingen benaderen het maximumniveau aan spellingvaardigheid dat gegeven de einddoelen te bereiken is.

In tabel 6.3b staat een overzicht van de vaardigheidsverdeling per normeringsmoment van Spelling werkwoorden. Hier nemen we voor de normeringsmomenten M7 t/m M8 de schatting van het gemiddelde, de standaarddeviatie en het aantal leerlingen in de normeringssteekproef op. M7 was immers het eerste normeringsmoment voor Spelling werkwoorden. Deze gegevens komen uit deze en uit onze eerdere wetenschappelijke verantwoordingen van de LVS-toetsen Spelling 3.0.

*Tabel 6.3b Overzicht van de vaardigheidsverdelingen per normeringsmoment Spelling werkwoorden*

| Normerings-moment | Aantal leerlingen | Aantal scholen | Gemiddelde vaardigheid | Standaard-deviatie |
|---|---|---|---|---|
| M7 | 540 | 27 | 131,7 | 23,7 |
| E7 | 2418 | 119 | 136,2 | 23,2 |
| M8 | 1560 | 87 | 142,7 | 23,4 |

Ook hier zien we dat de gemiddelde vaardigheid van de leerlingen van afname tot afname toeneemt, wat evidentie vormt voor de validiteit van de toetsen Spelling werkwoorden.

*Divergente validiteit: Correlatie scores LVS 3.0 met diverse toetsen leervorderingen*
Aan de scholen die hebben deelgenomen aan de normeringsonderzoeken M8 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 geautomatiseerde dataretourfunctie van het Computerprogramma LOVS zijn van de leerlingen van het normeringsonderzoek ook de scores op andere leervorderingentoetsen van Cito beschikbaar te maken.
Op basis van algemene cognitieve verschillen in aanleg (intelligentie) is er altijd sprake van een zekere samenhang tussen scores op leervorderingentoetsen voor verschillende vakgebieden. Hoe sterk deze samenhang is, hangt af van het vakgebied. De verwachting is dat we een sterke samenhang aantreffen (tussen 0,70 en 0,80) tussen Spelling niet-werkwoorden en Spelling werkwoorden. Het zijn immers beide aspecten van spellingvaardigheid. De correlatie zal niet groter zijn dan 0,80 omdat bij werkwoordspelling specifieke regels voor werkwoordspelling van toepassing zijn en leerlingen geen beroep kunnen doen op het woordbeeld. Dat is ook de reden dat we voor spelling niet-werkwoorden en spelling werkwoorden niet kunnen volstaan met één en dezelfde latente trek spellingvaardigheid.
We verwachten een redelijke samenhang tussen spellingvaardigheid en de vaardigheid in technisch lezen. Het zijn immers beide ‘technische’ vaardigheden binnen het taaldomein. Voor de overige vakgebieden verwachten we matige correlaties tussen toetsscores voor spellingvaardigheid enerzijds en alle (andere) toetsscores die voor hetzelfde meetmoment beschikbaar zijn anderzijds. Hoewel begrijpend lezen en woordenschat ook tot het taaldomein behoren, zijn deze vaardigheden meer semantisch van aard en daarmee duidelijk te onderscheiden van de vaardigheid spelling. Voor het spellen van woorden is een zekere basiswoordenschat weliswaar van belang, maar het is niet nodig om alle woordbetekenissen te kennen om woorden correct te kunnen schrijven. Voor een goede spellingvaardigheid zijn namelijk vooral de kennis en toepassing van regels en conventies van belang. Ook de correlatie met rekenen-wiskunde zal naar verwachting matig zijn, omdat dat een geheel andere vaardigheid betreft dan spelling.

In tabel 6.4 zijn de correlaties tussen de toets Spelling 3.0 M8 niet-werkwoorden en de toetsen Spelling werkwoorden M8, Technisch lezen (DMT) M8, Begrijpend lezen M8, Woordenschat M8 en Rekenen-Wiskunde M8 (alle van de tweede generatie) weergegeven.
In tabel 6.5 zijn de correlaties tussen de toets Spelling 3.0 M8 werkwoorden en de toetsen Spelling niet-werkwoorden M8, Technisch lezen (DMT) M8, Begrijpend lezen M8, Woordenschat M8 en Rekenen-Wiskunde M8 weergegeven (alle van de tweede generatie).

*Tabel 6.4 Correlaties tussen Spelling 3.0 M8 niet-werkwoorden en verschillende andere LVS-toetsen*

| | Spelling M8* | Aantal leerlingen |
|---|---|---|
| Cito Spelling werkwoorden M8 | 0,68 | 309 |
| Cito Technisch lezen – DMT M8 | 0,58 | 338 |
| Cito Begrijpend Lezen M8 | 0,55 | 321 |
| Cito Woordenschat M8 | 0,52 | 174 |
| Cito Rekenen-Wiskunde M8 | 0,45 | 332 |

* Deze correlaties zijn gecorrigeerd voor attenuatie.

*Tabel 6.5 Correlaties tussen Spelling 3.0 M8 werkwoorden en verschillende andere LVS-toetsen*

| | Spelling M8* | Aantal leerlingen |
|---|---|---|
| Cito Spelling niet-werkwoorden M8 | 0,71 | 312 |
| Cito Technisch lezen – DMT M8 | 0,36 | 307 |
| Cito Begrijpend Lezen M8 | 0,56 | 287 |
| Cito Woordenschat M8 | 0,58 | 143 |
| Cito Rekenen-Wiskunde M8 | 0,54 | 301 |

* Deze correlaties zijn gecorrigeerd voor attenuatie.

De samenhang tussen Spelling niet-werkwoorden en Spelling werkwoorden is niet hetzelfde voor beide normeringsonderzoeken op M8. In tabel 6.4 gaat het om de samenhang tussen de normeringstoets Spelling 3.0 M8 niet-werkwoorden en de toets Spelling M8 werkwoorden van de tweede generatie. In tabel 6.5 gaat het om de samenhang tussen de normeringstoets Spelling 3.0 M8 werkwoorden en de toets Spelling M8 niet-werkwoorden van de tweede generatie. In beide gevallen treffen we volgens verwachting een sterke samenhang aan (rond de 0,70).
Er is volgens verwachting een redelijke samenhang (r = 0,58) tussen spelling niet-werkwoorden en technisch lezen. Bij spelling werkwoorden valt op dat de correlatie met technisch lezen daar een stuk lager is dan bij spelling niet-werkwoorden (namelijk 0,36). Waarschijnlijk heeft dit ermee te maken dat er bij werkwoordspelling ook veel andere aspecten een rol spelen dan alleen het technische aspect, zoals het doorzien van grammaticale structuren.
Uit de tabellen 6.4 en 6.5 blijkt verder dat de correlaties tussen spelling en de andere vakgebieden (begrijpend lezen, woordenschat, rekenen-wiskunde) middelmatig hoog zijn. Dat is volgens verwachting: de spellingvaardigheid heeft een eigen structuur, die tot op grote hoogte bepaald wordt door kennis en toepassing van regels en conventies. De andere vaardigheden zijn meer semantisch van aard en daarmee duidelijk te onderscheiden van de vaardigheid spelling.

Alle correlaties met toetsen die andere vaardigheden dan spellingvaardigheid meten zijn aanzienlijk lager dan de (hoge) correlaties met soortgenoten (> 0,80) die we eerder in deze paragraaf rapporteerden.

Samenvattend kan dus gesteld worden dat de correlaties van de toetsen Spelling 3.0 met andere toetsen voor leervorderingen conform de verwachtingen zijn. Met het onderdeel passieve spelling uit de toets Taalverzorging is er overeenkomstig onze uitgangspunten een sterke samenhang. De correlaties van de nieuwe toetsen Spelling 3.0 niet-werkwoorden voor groep 8 met de toetsen Spelling van de tweede generatie, waarvan de begripsvaliditeit positief is beoordeeld, en met de toetsen Spelling 3.0 voor groep 7 zijn bovendien zeer hoog. Al deze correlaties zijn duidelijk hoger dan de correlaties met andere leervorderingentoetsen die geen spellingvaardigheid meten en vormen daarmee een ondersteuning voor de begripsvaliditeit van de toets in termen van convergente en divergente validiteit. De data geven aan dat er gemeten wordt wat men met de toetsen Spelling 3.0 beoog t te meten, namelijk spellingvaardigheid.

## 6.6 Verschillen tussen relevante subgroepen

Bij de normeringsonderzoeken M8 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 8 nog geen 11 jaar waren, reguliere leerlingen zijn leerlingen die op dat moment tussen de 11 en de 12 jaar waren en de groep oudere leerlingen bestaat uit leerlingen die op 1 oktober in groep 8 12 jaar of ouder waren. In tabel 6.6a en 6.6b wordt de gemiddelde score van de verschillende leeftijdsgroepen weergegeven. De effectgrootte wordt berekend ten opzichte van het overall-gemiddelde.

*Tabel 6.6a Gemiddelde score per leeftijdsgroep voor de afnamemoment M8 voor Spelling 3.0 niet-werkwoorden*

**M8**

| Groep | Aantal | M | SD | Effectgrootte (*d*) |
|---|---|---|---|---|
| Jonger | 101 | 382,9 | 30,1 | 0,50 |
| Regulier | 708 | 369,2 | 29,5 | 0,05 |
| Ouder | 119 | 347,2 | 23,9 | -0,70 |

*Tabel 6.6b Gemiddelde score per leeftijdsgroep voor de afnamemoment M8 voor Spelling 3.0 werkwoorden*

**M8**

| Groep | Aantal | M | SD | Effectgrootte (*d*) |
|---|---|---|---|---|
| Jonger | 81 | 157,0 | 29,8 | 0,43 |
| Regulier | 569 | 145,1 | 27,4 | 0,01 |
| Ouder | 82 | 130,1 | 24,8 | -0,53 |

Het patroon in tabel 6.6a en 6.6b is naar verwachting en komt overeen met eerder gevonden verschillen tussen reguliere en vertraagde, respectievelijk versnelde leerlingen (zie bijv. Van Til, Van Weerden, Hemker & Keune, 2014). Zowel bij Spelling niet-werkwoorden als bij Spelling werkwoorden scoort de groep jongere leerlingen hoger dan gemiddeld. Dit is een klein effect voor werkwoorden, maar voor niet-werkwoorden net een medium 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 bij beide toetsen het laagst. Bij zowel niet-werkwoorden als werkwoorden is sprake van een medium effect. Ook hier is dat naar verwachting, aangezien deze leerlingen in veel gevallen op grond van hun leerprestaties een jaar gedoubleerd hebben (of langer ‘gekleuterd’). De leerlingen die in de reguliere jaargroep zitten behalen een score die valt tussen de scores van de jongere en oudere leerlingen.

Bij de variabele geslacht is de gemiddelde score van jongens en meisjes berekend, zie tabel 6.7a en 6.7b.

*Tabel 6.7a Gemiddelde score jongen-meisje voor Spelling 3.0 niet-werkwoorden*

| Afname-moment | Geslacht | Aantal | M | SD |
|---|---|---|---|---|
| M8 | jongen | 460 | 366,3 | 31,3 |
| | meisje | 449 | 368,8 | 29,0 |

*Tabel 6.7b Gemiddelde score jongen-meisje voor Spelling 3.0 werkwoorden*

| Afname-moment | Geslacht | Aantal | M | SD |
|---|---|---|---|---|
| M8 | jongen | 358 | 141,3 | 28,3 |
| | meisje | 354 | 148,9 | 27,7 |

In tabel 6.7a en 6.7b zijn de gemiddelde scores van jongens en meisjes weergegeven voor afnamemoment M8. Het is te zien dat de gemiddelde scores van meisjes iets hoger liggen dan die van jongens. Alleen voor M8 werkwoorden is er sprake van een klein effect ($d$ = 0,27), voor M8 niet-werkwoorden niet.
Het is een bekend verschijnsel dat er bij spellingprestaties over het algemeen lichte verschillen zijn in het voordeel van de meisjes, die niet altijd substantieel zijn (zie bijv. Van Til, Van Weerden, Hemker & Keune, 2014) De uitkomsten passen bij deze verwachtingen: kleine verschillen, met een verwaarloosbaar of klein effect.

Ten slotte is er gekeken naar het leerlinggewicht. In tabel 6.8a en 6.8b zijn de gemiddelde scores 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 6.8a Gemiddelde score naar leerlinggewicht voor Spelling 3.0 groep 8 niet-werkwoorden*

| Moment | Gewicht | Aantal | M | SD | Effectgrootte ($d$) |
|---|---|---|---|---|---|
| M8 | 0,0 | 457 | 369,5 | 31,0 | 0,06 |
| | 0,3 | 37 | 360,7 | 26,8 | -0,24 |
| | 1,2 | 23 | 354,7 | 33,5 | -0,44 |

*Tabel 6.8b Gemiddelde score naar leerlinggewicht voor Spelling 3.0 groep 8 werkwoorden*

| Moment | Gewicht | Aantal | M | SD | Effectgrootte ($d$) |
|---|---|---|---|---|---|
| M8 | 0,0 | 363 | 147,8 | 29,5 | 0,11 |
| | 0,3 | 22 | 132,7 | 26,7 | -0,43 |
| | 1,2 | 15 | 135,1 | 25,3 | -0,34 |

In tabel 6.8a en 6.8b is te zien dat leerlingen met een leerlinggewicht van 0,3 en 1,2 op M8 niet-werkwoorden inderdaad lager scoren dan leerlingen zonder leerlinggewicht (gewicht van nul). In termen van de effectgrootte Cohens *d* is er sprake van een klein effect voor leerlingen met een gewicht bij spelling niet-werkwoorden. Voor Spelling M8 werkwoorden geldt hetzelfde. Ook deze bevindingen komen overeen met eerder gevonden verschillen: de leerlingen zonder leerlinggewicht scoren het hoogst, de leerlingen met een gewicht van 0,3 doen het duidelijk minder en de leerlingen met een gewicht van 1,2 scoren over het algemeen het laagst (Van Til, Van Weerden, Hemker & Keune, 2014). Hier moet opgemerkt worden dat dit niet het geval is voor M8 werkwoorden. Hier scoren de leerlingen met een hoger gewicht net iets beter dan leerlingen met een lager gewicht. Het gaat hier echter om zo'n kleine groep dat hier geen conclusies aan kunnen worden verbonden.

Waar wij op theoretische gronden verwachtingen hadden over verschillen tussen subgroepen, zijn deze bevestigd. Deze resultaten zijn op te vatten als aanvullend bewijsmateriaal ten aanzien van 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 8 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 lagere 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. Voor de toetsen Spelling in groep 8 gebruiken we twee opgavenbanken: één voor Spelling niet-werkwoorden en één voor Spelling werkwoorden. We lichtten toe dat de vaardigheid spelling niet-werkwoorden kan worden opgevat als een unidimensionaal continuüm en dat ditzelfde geldt voor de vaardigheid in werkwoordspelling. 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 categorieënoverzichten: één voor spelling niet-werkwoorden en één voor spelling werkwoorden. 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 de toetsitems succesvol konden worden geschaald en dat het functioneren van leerlingen op de toetsen terug te voeren is op één unidimensionaal concept (voor zowel spelling niet-werkwoorden als spelling werkwoorden afzonderlijk). In paragraaf 4.3.2 werd voorts aangetoond dat de normeringssteekproeven M8 op basis van de variabelen regio, urbanisatiegraad, schooltype en sekse een goede afspiegeling vormen van de populatie. De data in deze steekproeven vormen een combinatie van uitkomsten van toetsafnames in de vorm van *embedded field* onderzoek en dataretour. We betoogden dat de gekozen aanpak de best mogelijke garantie vormt voor een adequate initiële normering. In de laatste paragraaf van hoofdstuk 4 presenteerden we de normerings-resultaten 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 betrouwbaarheidscoëfficiënten van de toetsen zijn met 0,88 en 0,91 in relatie tot het beoogde gebruik van de toetsen zeer goed te noemen. Verder zijn in dit hoofdstuk betrouwbaarheidstabellen opgenomen die de betekenis van de meetnauw-keurigheid 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 toetsen Spelling niet-werkwoorden en Spelling werkwoorden, 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 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 zeer hoge correlatie van de LVS-toetsen Spelling 3.0 met de Cito LVS-toetsen Spelling van de tweede generatie aangevoerd (0,91 voor Spelling niet-werkwoorden M8 en 0,92 voor Spelling werkwoorden M8). De LVS-toetsen Spelling van de tweede generatie toetsen werden eerder door de COTAN op begrips-validiteit positief beoordeeld. We hebben ook de samenhangen met de toetsen Spelling 3.0 voor groep 7 kunnen berekenen. Deze zijn zeer hoog. Op basis hiervan is de convergente validiteit van de nieuwe toetsen hoog te noemen. De correlaties met andere toetsen op het gebied van leervorderingen bleken lager dan de correlatie tussen de toetsen Spelling onderling. Dit kan als bewijs van divergente validiteit worden opgevat. Voorts bleken de gemiddelde vaardigheidsscores op de normeringsmomenten voor de toetsen spelling in groep 8 uitstekend te passen in de reeks gemiddelden die eerder geconstrueerde en verantwoorde toetsen Spelling laten zien voor eerdere afnamemomenten. Er is sprake van stelselmatige, maar afvlakkende vaardigheidsgroei zoals die op theoretische gronden mocht worden verwacht.
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 8 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 (2018). *Cito Volgsysteem primair en speciaal onderwijs. Spelling 3.0 Groep 8*. 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.

Frith, U. (1980). Unexpected spelling problems. In U. Frith (Ed.), *Cognitive processes in spelling.* (pp. 495-515). London: Academic Press.

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. (1988). The Rasch model and multi-stage testing. *Journal of Educational Statistics, 13*, 45-52.

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 schoolniveau met Cito dataretour*. Arnhem: Cito.

Keuning, J., Boxtel, H. van, Lansink, N., Visser, J., Weekers, A. & Engelen, R. (2015). *Actualiteit en kwaliteit van normen. Een werkwijze voor het normeren van een leerlingvolgsysteem.* Arnhem: Cito.

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.

Tak, J.A., Bosch, J.D., Begeer, S. & Albrecht, G. (red.). (2014). *Handboek psychodiagnostiek voor de hulpverlening aan kinderen en adolescenten.* Utrecht: Uitgeverij De Tijdstroom.

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.

Tomesen, M., Wouda, J., Krämer, I. & Horsels, L. (2018). *Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 7*. Arnhem: Cito.

Verhaert, N. & D. Sandra (2016). Homofoondominantie veroorzaakt dt-fouten tijdens het spellen en maakt er ons blind voor tijdens het lezen. *Levende Talen Tijdschrift, 4*.

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.

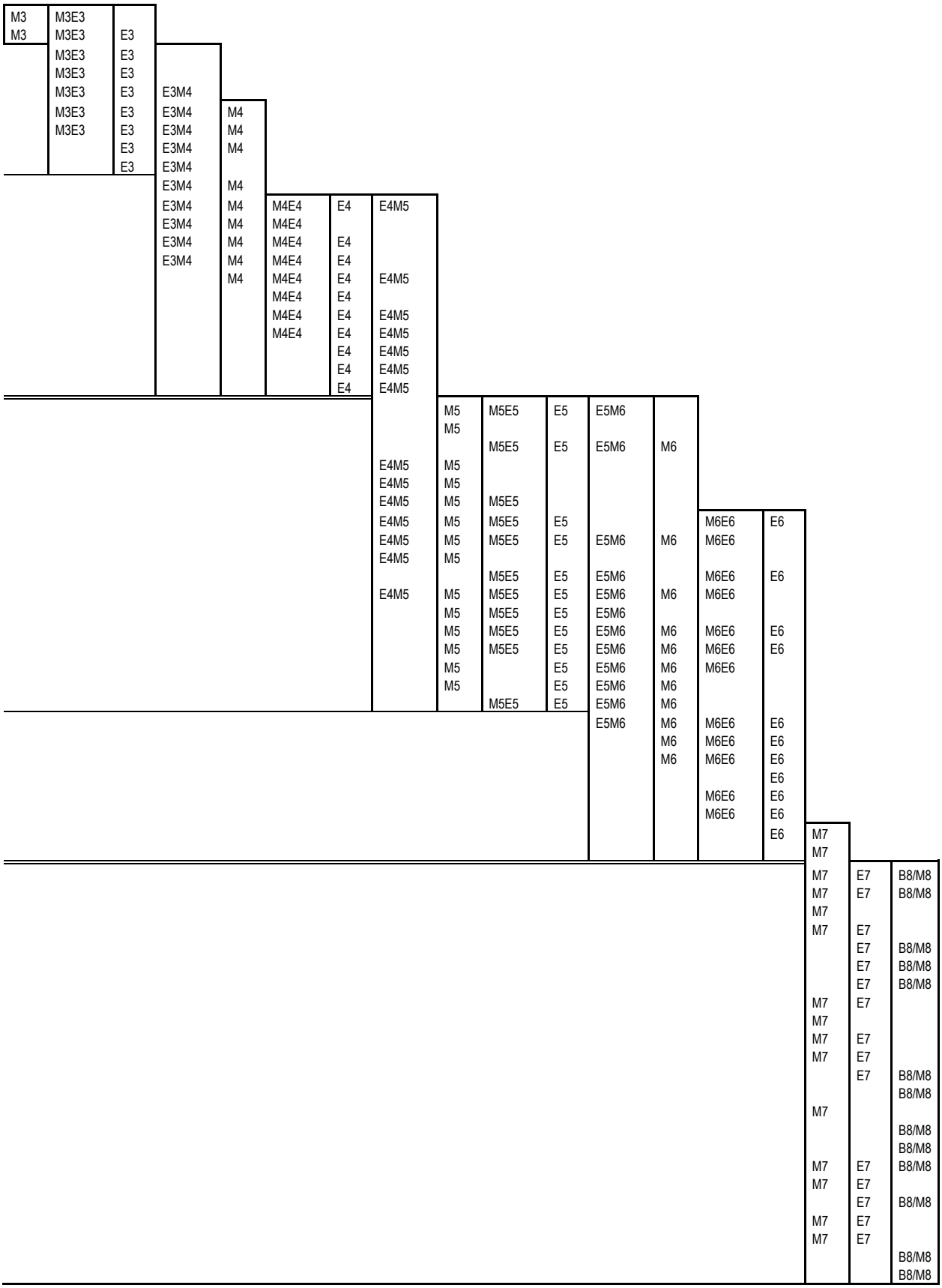
Verhelst, N.D., H.H.F.M. Verstralen & T.H.J.M. Eggen (1991). Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model. *Measurement and Research Department Reports 91-10*. Arnhem: Cito.

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

# Bijlagen

## Bijlage 1 Categorieënoverzicht Spelling niet-werkwoorden

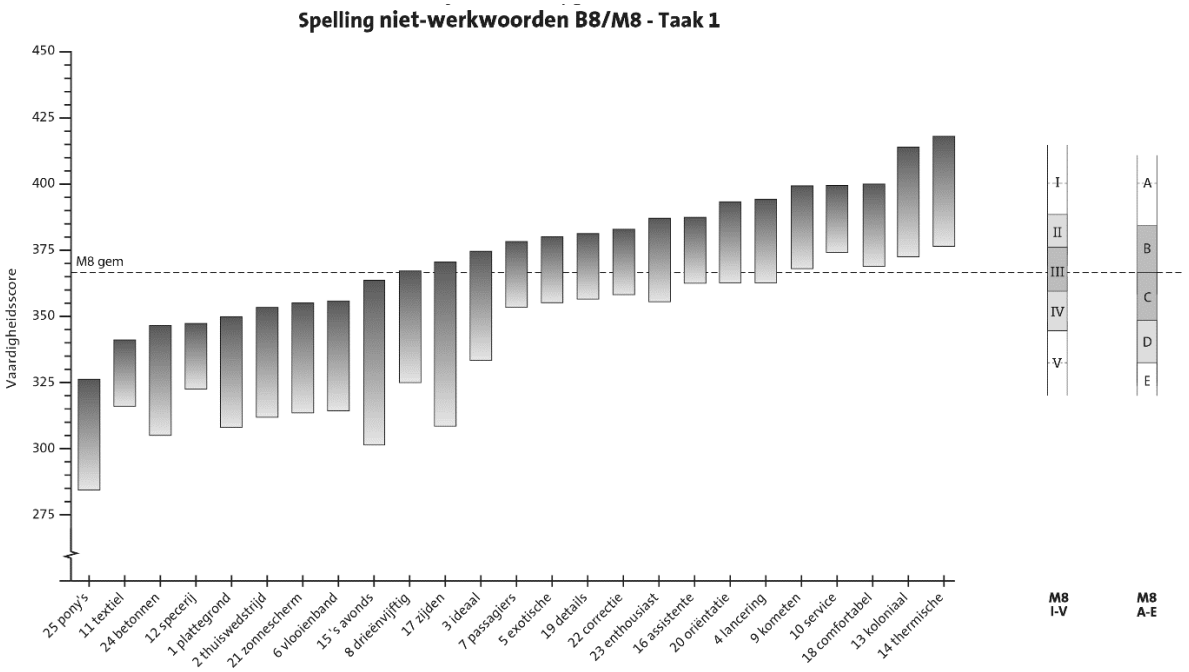
| cat | omschrijving | aantal lettergrepen | voorbeelden |
|---|---|---|---|
| 1 | woorden met mkm | een lettergreep | bal, pop, boom |
| 2 | woorden met mmkm en mkmm | een lettergreep | klap, bries, paars |
| 3 | woorden met mmkmm | een lettergreep | klomp, plant |
| 4 | woorden met een niet geschreven tussenklank | een lettergreep | erf, melk |
| 5 | woorden met -(m)mmm of mmm- | een lettergreep | kunst, strik |
| 6 | woorden met sch(r)- | een lettergreep | school, schrik |
| 7 | woorden met -ng(-) of -nk(-) | een lettergreep | slang, links |
| 8 | woorden met f- of v- en s- of z- | een of twee lettergrepen | feest, verf, ziek |
| 9 | verkleinwoord met uitgang -je(s) of -tje(s) | twee lettergrepen | huisje, stoeltje |
| 10 | woorden met -aai(-), -ooi(-), -oei(-) | een lettergreep | saai, mooi, boei |
| 11 | samengestelde woorden met twee of meer medeklinkers na elkaar | twee of meer lettergrepen | schooltas, fietstocht |
| 12 | woorden met -eer(-), -eur(-), -oor(-) | een of twee lettergrepen | beer, geur, poort |
| 13 | woorden met be-, ge-, ver-, te-, -el, -er, -en, -te | twee lettergrepen | begin, gevaar, diepte |
| 14 | woorden met (-)ei(-) of (-)ij(-) | een of twee lettergrepen | geit, fijne |
| 15 | woorden eindigend op -d | een of twee lettergrepen | hond, eend |
| 16 | woorden op -a, -o, -u (klinkt als /aa/, /oo/, /uu/) | een of twee lettergrepen | zo, nu |
| 17 | woorden met -au(-), -auw, -ou(-) of -ouw | een of twee lettergrepen | gauw, zout |
| 18 | woorden met -ch(t) | een of twee lettergrepen | pech, nacht |
| 19 | woorden op -eeuw, -ieuw, -uw | een of twee lettergrepen | leeuw, nieuw, ruw |
| 20 | woorden met open eerste lettergreep | twee lettergrepen | bomen, hamer |
| 21 | woorden met gesloten eerste lettergreep | twee lettergrepen | bruggen, petten |
| 7+ | woorden met -ng(-) of -nk(-) | twee of meer lettergrepen | lengte, donker |
| 8+ | woorden met (-)f- of (-)v- en (-)s- of (-)z- | twee of meer lettergrepen | zonder, ventiel, boven |
| 9+ | verkleinwoord met uitgang -je(s) (na -d en -t), -etje(s), -pje(s) | twee of meer lettergrepen | feestje, boompje, dingetje |
| 10+ | woorden met (-)aai(-), (-)ooi(-), -oei(-) | twee of meer lettergrepen | lawaai, rotzooi, geloei |
| 12+ | woorden met (-)eer(-), (-)eur(-), (-)oor(-) | twee of meer lettergrepen | onweer, monteur |
| 13+ | woorden met be-, ge-, ver-, te-, -el, -er, -en, -te | twee of meer lettergrepen | verzinsel, bedrieger |
| 14+ | woorden met (-)ei(-) of (-)ij(-) | twee of meer lettergrepen | refrein, aardbei, vijver |
| 15+ | woorden met -d(-) (klinkt als /t/) | twee of meer lettergrepen | handvat, rashond |
| 16+ | woorden op -a, -o, -u (klinkt als /aa/, /oo/, /uu/) | twee of meer lettergrepen | tempo, paraplu |
| 17+ | woorden met -au(-), -auw(-), -ou(-) of -ouw(-) | twee of meer lettergrepen | gemauw, kabouter |
| 18+ | woorden met -ch(t)(-) | twee of meer lettergrepen | echo, jochie, gewicht |
| 19+ | woorden met -eeuw(-), -ieuw(-) of -uw(-) | twee of meer lettergrepen | leeuwin, zwaluw |
| 20+ | woorden met open lettergreep | twee of meer lettergrepen | soldaten, bananen |
| 21+ | woorden met gesloten lettergreep | twee of meer lettergrepen | trommel, oppasser |
| 22 | verandering van -f in -v- en -s in -z- bij verbuiging of meervoudsvorming | twee of meer lettergrepen | brave, huizen |
| 23 | woorden met -em, -elen, -enen of -eren | twee of meer lettergrepen | kinderen, stiekem |
| 24 | woorden op -ig(-) en -lijk(-) | twee of meer lettergrepen | stevige, lelijk |
| 25 | woorden waarin /ie/ geschreven wordt als i | twee lettergrepen | prima, minuut, gitaar |
| 26 | woorden waarin /s/ geschreven wordt als c | een of meer lettergrepen | cel, december |
| 27 | woorden waarin /k/ geschreven wordt als c | een of meer lettergrepen | clown, camera |
| 28 | woorden beginnend met 's of eindigend op 's | een of meer lettergrepen | 's middags, zebra's |
| 29 | woorden met -tie(-) waarin t klinkt als ts | twee of meer lettergrepen | vakantie, traditie |
| 30 | woorden met -heid of -teit | twee of meer lettergrepen | kwaliteit, eenheid |
| 31 | leenwoorden waarin /zj/ geschreven wordt als g(e) | twee of meer lettergrepen | giraf, horloge |
| 32 | leenwoorden waarin /sj/ geschreven wordt als ch (nieuw) | een of meer lettergrepen | chef, machine |
| 20++ | woorden met open lettergreep | twee of meer lettergrepen | notaris, folie |
| 21++ | woorden met gesloten lettergreep | twee of meer lettergrepen | bemanning, terras |
| 24+ | woorden met -ig(-) of -lijk(-) | twee of meer lettergrepen | akelig, koninklijk |
| 25+ | woorden waarin /ie/ geschreven wordt als i | twee of meer lettergrepen | diploma, kiwi, televisie |
| 26+ | woorden waarin /s/ geschreven wordt als c | twee of meer lettergrepen | narcis, ceremonie |
| 27+ | woorden waarin /k/ geschreven wordt als c | twee of meer lettergrepen | camera, accordeon |
| 28+ | woorden beginnend met 's of eindigend op 's | twee of meer lettergrepen | 's ochtends, komma's |
| 29+ | woorden met -tie(-) waarin t klinkt als (t)s | twee of meer lettergrepen | vakantie, traditie |
| 30+ | woorden met -heid of -teit | twee of meer lettergrepen | majesteit, veiligheid |
| 33 | woorden met -b(-) | een of meer lettergrepen | krab, absoluut |
| 34 | woorden met (-)y(-) | twee of meer lettergrepen | pony, baby, pyjama |
| 35 | woorden met een trema | twee of meer lettergrepen | zeeën, ruïne |
| 37 | samenstelling met tussen -e(n)- | twee of meer lettergrepen | blokkendoos, zonnehoed |
| 38 | woorden met of zonder een hoofdletter | een of meer lettergrepen | Marjolein, Pasen, kerst, juni |
| 39 | Franse leenwoorden | een of meer lettergrepen | bureau, journaal |
| 40 | Engelse leenwoorden | een of meer lettergrepen | team, keeper |
| 41 | woorden waarin /t/ geschreven wordt als th | een of meer lettergrepen | theater, thuis, apotheek |
| 42 | woorden met -isch(e) | twee of meer lettergrepen | alfabetisch, automatisch |
| 43 | woorden waarin /ks/ geschreven wordt als x | twee of meer lettergrepen | taxi, examen |
| 44 | verkleinwoorden -aatje, -eetje, -ootje, -uutje en met de uitgang -nkje | twee of meer lettergrepen | parapluutje, kettinkje |
| 45 | woorden met assimilatieverschijnselen | twee of meer lettergrepen | zakdoek,ontdekking |
| 46 | woorden op -iaal, -ieel, -ueel, -eaal | meer lettergrepen | speciaal, commercieel, actueel |
| 47 | stoffelijke bijvoeglijke naamwoorden | twee of meer lettergrepen | wollen, leren, houten |

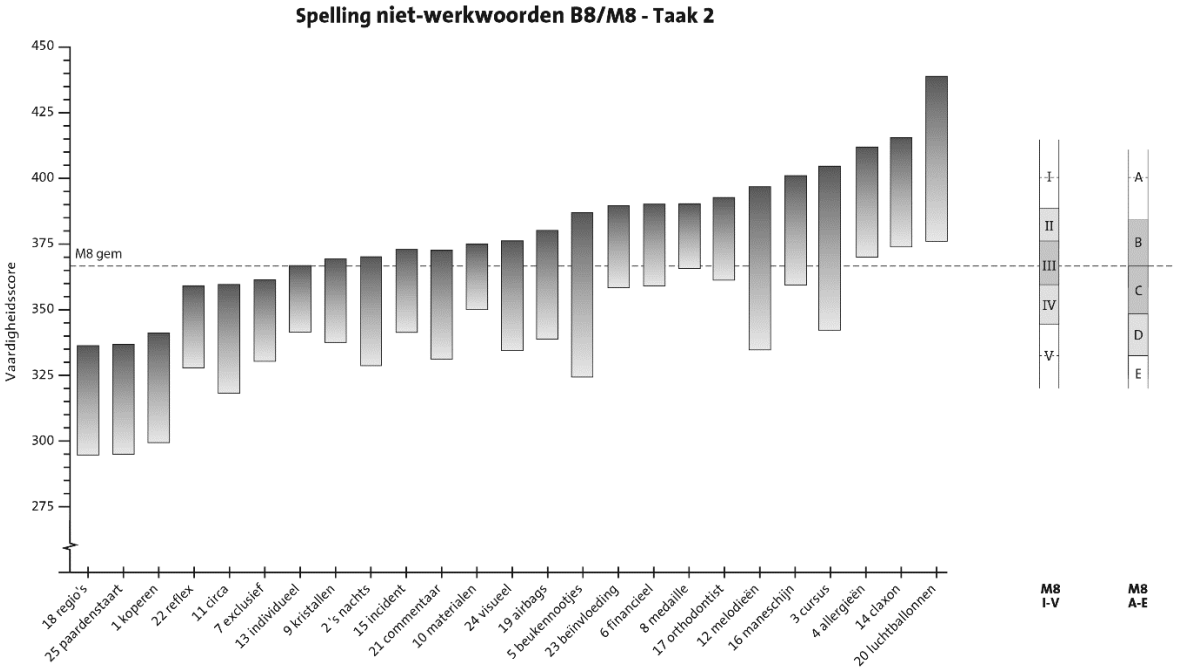

M3 M3E3
M3 M3E3 E3
M3E3 E3
M3E3 E3
M3E3 E3 E3M4
M3E3 E3 E3M4 M4
M3E3 E3 E3M4 M4
E3 E3M4 M4
E3 E3M4
E3M4 M4
E3M4 M4 M4E4 E4 E4M5
E3M4 M4 M4E4
E3M4 M4 M4E4 E4
E3M4 M4 M4E4 E4
M4 M4E4 E4 E4M5
M4E4 E4
M4E4 E4 E4M5
M4E4 E4 E4M5
E4 E4M5
E4 E4M5
E4 E4M5
M5 M5E5 E5 E5M6
M5
M5E5 E5 E5M6 M6
E4M5 M5
E4M5 M5
E4M5 M5 M5E5
E4M5 M5 M5E5 E5 M6E6 E6
E4M5 M5 M5E5 E5 E5M6 M6 M6E6
E4M5 M5
M5E5 E5 E5M6 M6E6 E6
E4M5 M5 M5E5 E5 E5M6 M6 M6E6
M5 M5E5 E5 E5M6
M5 M5E5 E5 E5M6 M6 M6E6 E6
M5 M5E5 E5 E5M6 M6 M6E6 E6
M5 E5 E5M6 M6 M6E6
M5 E5 E5M6 M6
M5E5 E5 E5M6 M6
E5M6 M6 M6E6 E6
M6 M6E6 E6
M6 M6E6 E6
E6
M6E6 E6
M6E6 E6
E6 M7
M7
M7 E7 B8/M8
M7 E7 B8/M8
M7
M7 E7
E7 B8/M8
E7 B8/M8
E7 B8/M8
M7 E7
M7
M7 E7
M7 E7
E7 B8/M8
B8/M8
M7
B8/M8
B8/M8
M7 E7 B8/M8
M7 E7
E7 B8/M8
M7 E7
M7 E7
B8/M8
B8/M8

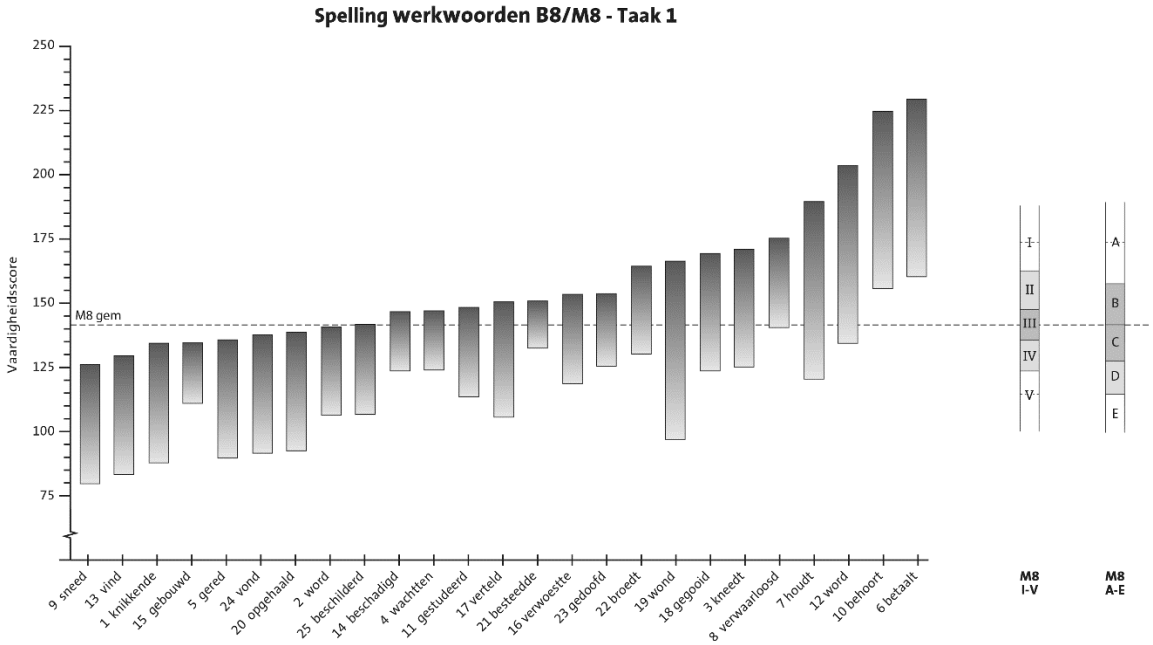
## Bijlage 2 Spellingcategorieën werkwoordspelling

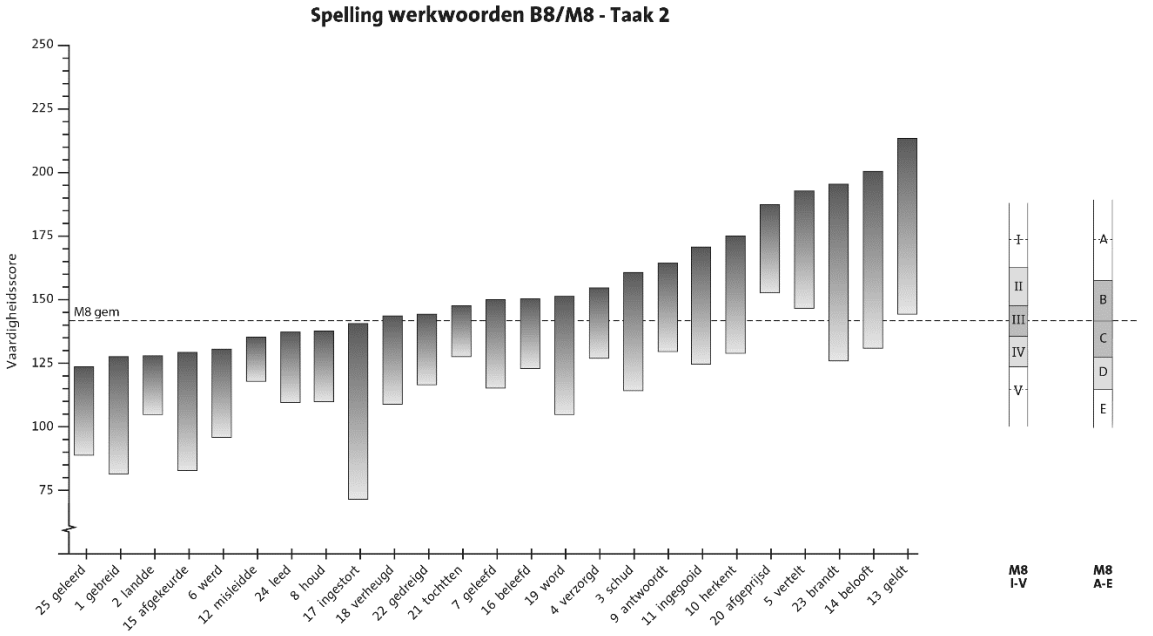
| Cat. | Omschrijving | Subcategorie | | Persoon | Voorbeelden | M7 | E7 | B8/M8 |
|---|---|---|---|---|---|---|---|---|
| **1** | **tijd van nu (o.t.t.)** | 1.a | -t achter stam van zwak ww dat in o.v.t. de uitgang -de krijgt | 23ev | jij tekent, zij geeft | X | X | |
| | | 1.b | wel of geen -t achter een stam op -d | 123ev | ik vind, jij onthoudt, zij wordt, hij verbindt | X | X | X |
| | | 1.c | bij inversie pv-ond: wel of geen -t achter een stam op -d (vraag of gebiedende wijs) | 123ev | bind ik? word jij? houdt u? schud de kaarten! | | | X |
| | | 1.d | homofone gevallen | 23ev | het gebeurt, jij verdeelt | | | X |
| **2** | **tijd van toen (o.v.t.)** | 2.a | zwak ww dat in o.v.t. de uitgang -te(n) of -de(n) krijgt | 123ev 123mv | ik bakte, zij tekende, wij hoopten | X | X | |
| | | 2.b | verdubbeling d of t bij zwak ww met stam op -d of -t | 123ev 123mv | ik raadde, jij stootte, wij landden | X | X | X |
| | | 2.c | geen -t bij sterk ww dat in 2e en 3e persoon eindigt op -d | 23ev | jij werd, zij hield, hij stond | X | X | X |
| | | 2.d | uitgang -sde(n) of -fde(n) bij zwak ww met stam op -z of -v | 123ev 123mv | ik beefde, de storm raasde, jullie verfden | X | X | |
| **3** | **voltooid deelwoord** | 3.a | keuze voor eind-d of eind-t bij zwakke werkwoorden met een stam die **niet** eindigt op -d, -t, -v of -z | - | geblust, gezeurd | | X | X |
| | | 3.b | homofone gevallen | - | is beoordeeld, is verbrand | | X | X |
| | | 3.c | zwakke werkwoorden met stam op -d, -t, -v of -z | - | gebeefd, geraasd, afgemeld | | X | X |
| **4** | **(on)voltooid deelwoord bijvoeglijk gebruikt** | 4.a | wel of geen -n aan het eind; -d of -t aan het eind; onvoltooid deelwoord bijvoeglijk gebruikt | - | gekookte eieren, gebraden vlees; gegrild vlees, ingeblikt voedsel; trillende stem | | X | X |

## Bijlage 3 Moeilijkheid van opgaven per taak in Spelling 3.0 groep 8




Spelling werkwoorden B8/M8 - Taak 1
Vaardigheidsscore
250
225
200
175
150
125
100
75
M8 gem
9 sneed
13 vind
1 knikkende
15 gebouwd
5 gered
24 vond
20 opgehaald
2 word
25 beschilderd
14 beschadigd
4 wachtten
11 gestudeerd
17 verteld
21 besteedde
16 verwoeste
23 gedoofd
22 broedt
19 wond
18 gegooid
3 kneedt
8 verwaarloosd
7 houdt
12 word
10 behoort
6 betaalt
I
II
III
IV
V
A
B
C
D
E
M8 I-V
M8 A-E


Spelling werkwoorden B8/M8 - Taak 2
Vaardigheidsscore
250
225
200
175
150
125
100
75
M8 gem
25 geleerd
1 gebreid
2 landde
15 afgekeurde
6 werd
12 misleidde
24 leed
8 houd
17 ingestort
18 verheugd
22 gedreigd
21 tochtten
7 geleefd
16 beleefd
19 word
4 verzorgd
3 schud
9 antwoordt
11 ingegooid
10 herkent
20 afgeprijsd
5 vertelt
23 brandt
14 belooft
13 geldt
I
II
III
IV
V
A
B
C
D
E
M8 I-V
M8 A-E

## Bijlage 4 Klassieke en IRT-indices van de opgaven in toetsen Spelling 3.0 groep 8 niet-werkwoorden en werkwoorden

*Toets B8/M8 niet-werkwoorden*

Normeringsmoment M8

| Volgnr | P-Val | RIT | Beta | Info |
|---|---|---|---|---|
| 1 | 0,845 | 0,335 | -0,047 | 1,055 |
| 2 | 0,830 | 0,345 | -0,006 | 1,130 |
| 3 | 0,724 | 0,394 | 0,231 | 1,551 |
| 4 | 0,538 | 0,498 | 0,560 | 3,047 |
| 5 | 0,616 | 0,556 | 0,474 | 4,140 |
| 6 | 0,819 | 0,352 | 0,023 | 1,183 |
| 7 | 0,629 | 0,555 | 0,457 | 4,090 |
| 8 | 0,767 | 0,378 | 0,143 | 1,402 |
| 9 | 0,493 | 0,495 | 0,619 | 3,061 |
| 10 | 0,431 | 0,542 | 0,689 | 4,266 |
| 11 | 0,887 | 0,422 | 0,042 | 1,975 |
| 12 | 0,856 | 0,455 | 0,113 | 2,367 |
| 13 | 0,464 | 0,415 | 0,666 | 1,887 |
| 14 | 0,437 | 0,411 | 0,709 | 1,869 |
| 15 | 0,795 | 0,271 | -0,122 | 0,614 |
| 16 | 0,548 | 0,558 | 0,554 | 4,302 |
| 17 | 0,770 | 0,280 | -0,044 | 0,664 |
| 18 | 0,488 | 0,494 | 0,625 | 3,060 |
| 19 | 0,604 | 0,557 | 0,487 | 4,177 |
| 20 | 0,542 | 0,498 | 0,554 | 3,043 |
| 21 | 0,822 | 0,350 | 0,015 | 1,169 |
| 22 | 0,588 | 0,558 | 0,507 | 4,224 |
| 23 | 0,597 | 0,497 | 0,481 | 2,962 |
| 24 | 0,857 | 0,325 | -0,084 | 0,990 |
| 25 | 0,918 | 0,265 | -0,308 | 0,632 |
| 26 | 0,876 | 0,310 | -0,144 | 0,886 |
| 27 | 0,750 | 0,385 | 0,178 | 1,463 |
| 28 | 0,625 | 0,313 | 0,331 | 0,864 |
| 29 | 0,480 | 0,416 | 0,640 | 1,893 |
| 30 | 0,706 | 0,299 | 0,133 | 0,770 |
| 31 | 0,572 | 0,498 | 0,514 | 3,007 |
| 32 | 0,783 | 0,448 | 0,196 | 2,186 |
| 33 | 0,517 | 0,556 | 0,589 | 4,331 |
| 34 | 0,734 | 0,469 | 0,282 | 2,473 |
| 35 | 0,659 | 0,550 | 0,420 | 3,962 |
| 36 | 0,803 | 0,361 | 0,063 | 1,256 |
| 37 | 0,662 | 0,308 | 0,243 | 0,827 |
| 38 | 0,731 | 0,531 | 0,325 | 3,541 |
| 39 | 0,455 | 0,414 | 0,681 | 1,882 |
| 40 | 0,708 | 0,477 | 0,322 | 2,597 |
| 41 | 0,554 | 0,419 | 0,523 | 1,876 |
| 42 | 0,549 | 0,498 | 0,544 | 3,036 |
| 43 | 0,890 | 0,296 | -0,194 | 0,803 |
| 44 | 0,692 | 0,403 | 0,291 | 1,643 |
| 45 | 0,455 | 0,316 | 0,708 | 0,911 |
| 46 | 0,736 | 0,390 | 0,206 | 1,511 |
| 47 | 0,797 | 0,440 | 0,170 | 2,094 |
| 48 | 0,574 | 0,498 | 0,512 | 3,005 |
| 49 | 0,717 | 0,396 | 0,244 | 1,572 |
| 50 | 0,890 | 0,297 | -0,192 | 0,807 |

*Toets B8/M8 werkwoorden*

Normeringsmoment M8

| Volgnr | P-Val | RIT | Beta | Info |
|---|---|---|---|---|
| 1 | 0,821 | 0,297 | -0,122 | 1,221 |
| 2 | 0,783 | 0,387 | 0,061 | 2,352 |
| 3 | 0,624 | 0,354 | 0,251 | 1,889 |
| 4 | 0,705 | 0,523 | 0,239 | 5,475 |
| 5 | 0,815 | 0,301 | -0,107 | 1,252 |
| 6 | 0,425 | 0,266 | 0,602 | 0,925 |
| 7 | 0,610 | 0,268 | 0,204 | 0,901 |
| 8 | 0,530 | 0,429 | 0,405 | 3,307 |
| 9 | 0,852 | 0,280 | -0,202 | 1,061 |
| 10 | 0,446 | 0,268 | 0,556 | 0,935 |
| 11 | 0,735 | 0,405 | 0,136 | 2,654 |
| 12 | 0,547 | 0,271 | 0,342 | 0,938 |
| 13 | 0,840 | 0,287 | -0,169 | 1,126 |
| 14 | 0,709 | 0,522 | 0,235 | 5,442 |
| 15 | 0,810 | 0,481 | 0,110 | 4,233 |
| 16 | 0,702 | 0,414 | 0,184 | 2,830 |
| 17 | 0,738 | 0,331 | 0,056 | 1,580 |
| 18 | 0,635 | 0,353 | 0,234 | 1,869 |
| 19 | 0,710 | 0,253 | -0,030 | 0,786 |
| 20 | 0,802 | 0,307 | -0,076 | 1,317 |
| 21 | 0,635 | 0,574 | 0,321 | 7,569 |
| 22 | 0,615 | 0,428 | 0,300 | 3,162 |
| 23 | 0,666 | 0,482 | 0,262 | 4,340 |
| 24 | 0,805 | 0,306 | -0,084 | 1,300 |
| 25 | 0,778 | 0,389 | 0,068 | 2,384 |
| 26 | 0,846 | 0,283 | -0,186 | 1,093 |
| 27 | 0,850 | 0,453 | 0,049 | 3,596 |
| 28 | 0,689 | 0,344 | 0,144 | 1,737 |
| 29 | 0,660 | 0,483 | 0,270 | 4,378 |
| 30 | 0,484 | 0,356 | 0,465 | 2,002 |
| 31 | 0,837 | 0,357 | -0,042 | 1,922 |
| 32 | 0,724 | 0,408 | 0,153 | 2,719 |
| 33 | 0,791 | 0,442 | 0,101 | 3,349 |
| 34 | 0,618 | 0,428 | 0,296 | 3,152 |
| 35 | 0,601 | 0,356 | 0,288 | 1,928 |
| 36 | 0,628 | 0,353 | 0,245 | 1,882 |
| 37 | 0,788 | 0,566 | 0,179 | 6,997 |
| 38 | 0,499 | 0,271 | 0,442 | 0,945 |
| 39 | 0,561 | 0,271 | 0,311 | 0,932 |
| 40 | 0,840 | 0,287 | -0,170 | 1,124 |
| 41 | 0,695 | 0,476 | 0,228 | 4,164 |
| 42 | 0,799 | 0,227 | -0,286 | 0,617 |
| 43 | 0,766 | 0,394 | 0,088 | 2,466 |
| 44 | 0,741 | 0,330 | 0,050 | 1,569 |
| 45 | 0,429 | 0,418 | 0,528 | 3,258 |
| 46 | 0,682 | 0,568 | 0,277 | 7,162 |
| 47 | 0,746 | 0,462 | 0,165 | 3,780 |
| 48 | 0,585 | 0,270 | 0,260 | 0,919 |
| 49 | 0,794 | 0,441 | 0,096 | 3,317 |
| 50 | 0,868 | 0,333 | -0,111 | 1,639 |

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