Wetenschappelijke verantwoording Spelling 3.0 digitaal voor groep 8

Aanvulling bij de wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8

Marieke Tomesen, Jasper Wouda en Linda Horsels

cito.nl

Wetenschappelijke verantwoording Spelling 3.0 digitaal voor groep 8

Aanvulling bij de wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8

Cito | Primair en speciaal onderwijs

Marieke Tomesen
Jasper Wouda
Linda Horsels

Cito
Arnhem, 2020

Inhoud

1 Inleiding

Deze *Aanvulling bij de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8* heeft uitsluitend betrekking op de *digitale* toetsen Spelling 3.0 voor groep 8.
De inhoud van deze digitale toetsen komt grotendeels overeen met de inhoud van de papieren toetsen voor groep 8. Vandaar dat we voor de inhoudelijke aspecten grotendeels verwijzen naar de oorspronkelijke Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019). Op punten waar de digitale toetsen afwijken van de papieren toetsen, gaan we in deze aanvulling in.

Tezamen met de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019) en de inhoud van het (digitale) toetspakket Spelling groep 8 (Cito, 2018) levert deze aanvulling alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit van de digitale toetsen Spelling 3.0 groep 8. Het genoemde materiaal maakt een beoordeling van de digitale toetsen Spelling 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 Aanvulling bij de Wetenschappelijke verantwoording gaat alleen in op de normering van M8. T.z.t. verschijnt een addendum over B8.
– betrouwbaarheid
– validiteit

Het laatstgenoemde aspect betreft alleen begripsvaliditeit en geen criteriumvaliditeit. Omdat de toetsen van het LVS niet bedoeld zijn voor 'voorspellend gebruik' is criteriumvaliditeit niet van toepassing.
Deze aanvulling heeft met name betrekking op de normen (hoofdstuk 4) en de betrouwbaarheid en meetnauwkeurigheid (hoofdstuk 5). Voor de uitgangspunten van de toetsconstructie (hoofdstuk 2 en 3) en de begripsvaliditeit (hoofdstuk 6) van de toetsen Spelling 3.0 verwijzen we naar de oorspronkelijke Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019). De kwaliteit van het toetsmateriaal en van de handleiding is te bepalen door kennis te nemen van de inhoud van de (digitale) toetspakketten.

Waar in deze Wetenschappelijke verantwoording met betrekking tot de leerkracht en/of leerling ‘hij’ staat, kan uiteraard ook ‘zij’ worden gelezen.

2 Uitgangspunten van de toetsconstructie

Voor de uitgangspunten van de toetsconstructie verwijzen we naar de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019). Alles wat in hoofdstuk 2 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 wordt gezegd over de meetpretentie, het gebruiksdoel en de functie van de toetsen, is ook van toepassing op de digitale toetsen.
Meetpretentie, gebruiksdoel en functie zijn identiek voor de papieren en digitale toetsen. De papieren en digitale toetsen zijn immers op dezelfde manier opgebouwd en in grote lijnen gelijk aan elkaar. Voor zowel de papieren als de digitale toetsen geldt het volgende:

- Ze meten de actieve spelling doordat de leerling woorden moet opschrijven cq. intypen.
- Ze zijn bestemd voor leerlingen in groep 8 van het basisonderwijs en voor leerlingen in het speciaal basisonderwijs en in het speciaal onderwijs cluster 1, cluster 2 (leerlingen met een taalontwikkelingsstoornis) en cluster 4.
- Ze kunnen ook gebruikt worden voor leerlingen in andere leerjaren die werken op het niveau van groep 8.
- De toetsen zijn niet geschikt voor leerlingen met een (tijdelijk) beperkt gehoor.
- De toetsen hebben twee doelen: niveaubepaling en progressiebepaling.
- De gemaakte fouten kunnen geanalyseerd worden met het oog op het aanbieden van gerichte remediëring.

Van alle papieren toetsen Spelling 3.0 zijn ook digitale varianten beschikbaar. Dit betekent dat er voor groep 8 een digitale toets B8/M8 niet-werkwoorden en een digitale toets B8/M8 werkwoorden beschikbaar is. De papieren en digitale toetsen van een bepaald afnamemoment zijn uitwisselbaar. Dat betekent dat de leerkracht op een afnamemoment zelf kan kiezen of hij de leerling een papieren of digitale toets laat maken. De keuze heeft geen invloed op het volgen van de ontwikkeling van de spellingvaardigheid. De scores op de papieren en digitale toetsen zijn namelijk uitwisselbaar. Ze zijn echter niet identiek, wat betekent dat eenzelfde aantal ‘goed’ tot een andere vaardigheidsscore leidt. De omzetting van vaardigheidsscore naar niveau is echter hetzelfde.

De theoretische inkadering van de toetsen - zowel inhoudelijk als psychometrisch (zie paragraaf 2.4 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019) - geldt zowel voor de papieren toetsen als de digitale toetsen.

3 Beschrijving van de toetsen

3.1 Opbouw en structuur van de toetsen

Het digitale toetspakket Spelling 3.0 voor groep 8 uit het Cito Volsysteem primair en speciaal onderwijs bevat – net als de papieren toetsen voor groep 8 – in totaal twee toetsen: één toets Spelling niet-werkwoorden en één toets Spelling werkwoorden. Beide toetsen kunnen op twee afnamemomenten worden afgenomen: aan het begin van leerjaar 8 of halverwege leerjaar 8. De leerkracht bepaalt zelf op welk moment hij de toets afneemt.

De digitale varianten van de toetsen bevatten net als de papieren varianten twee taken van 25 opgaven en zijn op dezelfde manier opgebouwd. Het opgaventype (zinsdictee) is hetzelfde. Ook het toetsen op maat is evengoed mogelijk bij de digitale toetsen. We verwijzen daarom hier naar hoofdstuk 3 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019) voor een nadere toelichting.
De opgaven van de papieren en de digitale toetsen representeren dezelfde spellingcategorieën.
De dicteewoorden die in een digitale toets voorkomen kunnen echter verschillen van de papieren toets van hetzelfde niveau. De papieren en digitale toetsen zijn dus niet identiek, maar hebben wel een grote overlap.

De afname, scoring en verwerking van de resultaten verschilt van die van de papieren toetsen. We lichten ze hieronder toe.

Afname
De digitale toetsen worden individueel op de computer, laptop of chromebook gemaakt. Afhankelijk van het aantal beschikbare devices kunnen meerdere leerlingen gelijktijdig aan dezelfde toets werken. De leerling krijgt voorafgaand aan de toets een uitleg over het maken van de digitale opgaven en maakt een of twee oefenopgaven. Op die manier raakt de leerling vertrouwd met het type opgave.
Bij het maken van de instructie is rekening gehouden met speciale leerlingen; zo is de instructie bijvoorbeeld kort en zijn samengestelde zinnen zo veel mogelijk vermeden. Dergelijke uitgangspunten gaan niet ten koste van de reguliere leerlingen.
Daarna maakt de leerling de toetsopgaven. Bij de toets Spelling niet-werkwoorden ziet de leerling op het scherm een antwoordvak. De computer leest de opgave voor. Daarna verschijnt de cursor in het antwoordvak, zodat de leerling het dicteewoord kan intypen. Bij de toets Spelling werkwoorden ziet de leerling het hele werkwoord, en daaronder de zin met een invulvak. Het is niet mogelijk om tijdens het afspelen van de audio al een antwoord in te typen. De leerling kan, indien gewenst, de opgave nog een keer afspelen door op de ‘play-knop’ te klikken. Ook kan hij het volume aanpassen met de knop ‘luidspreker’; deze staat rechts boven in het scherm. Naast deze knop staat de knop voor het overzichtsscherm. Zo ziet de leerling in één oogopslag welke opgaven hij al gemaakt heeft, wat hij heeft ingevuld én welke opgaven hij nog moet maken. De leerling kan zijn antwoord aanpassen door de knop ‘backspace’ op het toetsenbord te gebruiken. Bij de toets B8/M8 niet-werkwoorden staan rechts in beeld zeven knoppen met letters met een leesteken: è, é, ë, ï, ü,’s en s’. De leerling kan er ook voor kiezen om het toetsenbord te gebruiken om letters met een leesteken op te nemen.
Als de leerling een antwoord heeft ingetypt, klikt hij op de knop ‘verder’. De gemaakte opgave krijgt dan een andere kleur (lichtgrijs). Het is mogelijk om via de navigatiebalk door de toets te navigeren en de opgaven in een andere volgorde te maken. Aan het einde van de toets kan de leerling zijn antwoorden controleren in het zogenaamde ‘overzichtsscherm’. Indien gewenst kan de leerling hier op een antwoord klikken en dit aanpassen. Pas als alle opgaven gemaakt zijn, kan hij de toets inleveren.

In de toetsmap Spelling 3.0 groep 8 (Cito, 2018) is een inhoudelijke handleiding opgenomen behorend bij de papieren en digitale toetsen. Hierin staan de uitgebreide afname-instructies voor de leerkracht.

Daarnaast is er een technische digitale handleiding voor alle leerjaren (Cito, 2019), die door scholen via Cito Portal is te downloaden.
Bij de digitale versies van de toetsen worden de antwoorden van de leerlingen door de computer gescoord en hoeft de leerkracht de toetsen niet zelf na te kijken. De leerkracht kan ervoor kiezen om een foutenanalyse uit te draaien waarin hij kan zien in welke spellingcategorieën de leerling veel fouten maakt.

Het maakt voor de resultaten niet uit of leerlingen de papieren of de digitale toetsen maken. De opgaven uit de papieren en de digitale toetsen liggen op één vaardigheidsschaal, waardoor de toetsresultaten onderling uitwisselbaar zijn. Bij de keuze voor de afname van ofwel de papieren ofwel de digitale toets kunnen verschillende overwegingen een rol spelen. Dit zijn overwegingen van zowel praktische aard (bijvoorbeeld de aanwezigheid van voldoende devices) als van meer inhoudelijke aard. Vooral voor leerlingen met concentratieproblemen, leerlingen die langzamer of juist veel sneller dan gemiddeld werken en leerlingen die afwezig waren bij de klassikale afname, kan een individuele, digitale afname prettig zijn. De leerling moet wel voldoende computervaardig zijn om woorden te kunnen intypen bij de digitale toets.

Scoring
De digitale toetsen worden geautomatiseerd nagekeken. De toetsscore wordt automatisch omgezet naar de bijbehorende vaardigheidsscore met een score-interval ofwel betrouwbaarheidsinterval.

Verwerking resultaten
Met het Computerprogramma LOVS kunnen diverse rapportages, zoals leerlingrapporten en groepsoverzichten, en een foutenanalyse worden opgevraagd.

3.2 Inhoudsverantwoording

3.2.1 Domeinbeschrijving en uitwerking in spellingcategorieën

Aan de digitale toetsen ligt dezelfde domeinbeschrijving en uitwerking in spellingcategorieën ten grondslag als aan de papieren toetsen. Hiervoor verwijzen we naar paragraaf 3.2.1 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019). De voor groep 8 afzonderlijk uitgewerkte overzichten van spellingcategorieën voor niet-werkwoorden cq. werkwoorden (zie tabel 3.1a cq. 3.1b) vormde de basis voor de itemconstructie en de selectie van items in de definitieve toetsen.

3.2.2 Itemconstructie, onderzoeken en selectie van opgaven

Itemconstructie
Er heeft geen speciale itemconstructie plaatsgevonden voor de digitale toetsen. Bij de digitale toetsen putten we uit de items van de itembank die is gevormd bij de constructie van de papieren toetsen.
De papieren opgaven werden ‘omgebouwd’ tot een digitaal afneembare versie. De instructies en dicteeopgaven zijn ingesproken als een voice-over aan de hand van scripts. Een toetsdeskundige was aanwezig bij de opnames om de gesproken teksten direct te beoordelen en waar nodig bij te sturen. Voorafgaand aan de opnames bespraken de ‘voice-over’ en de toetsdeskundige aan de hand van voorbeeldaudio het spreektempo. Bij twijfel over de uitspraak van een dicteewoord werd het uitspraakwoordenboek geraadpleegd (zie www.woorden.org).

Samenstelling definitieve toetsen
In januari 2018 vonden kalibratieonderzoeken plaats voor de digitale items (zie hoofdstuk 4). Bij de analyses is de kwaliteit van de afzonderlijke items en de totale verzameling voor afnamemoment M8 in kaart gebracht. Itemparameters en discriminatieparameters zijn geschat. Bij de analyses van de antwoorden van de leerlingen op de opgaven is nagegaan of de papieren items en de digitale items dezelfde vaardigheid meten en op dezelfde schaal passen. Dat bleek het geval te zijn.

Zie voor een uitgebreide verantwoording hoofdstuk 4. Voor de digitale items zijn eigen moeilijkheids- en discriminatieparameters geschat. Het is immers niet noodzakelijkerwijs zo dat de papieren versie en digitale versie van een item precies even moeilijk zijn en/of evengoed discrimineren. Met andere woorden: het is zeer wel mogelijk dat de papieren en digitale versie van hetzelfde item in deze kalibratie verschillende itemparameters krijgen toegekend.

Voor het samenstellen van de definitieve digitale toetsen zijn de volgende uitgangspunten gehanteerd:

- De digitale toetsen bevatten bij voorkeur precies dezelfde aantallen opgaven per categorie als de papieren toetsen.
- Van elke categorie zijn minimaal drie opgaven in een toets opgenomen.
- De digitale toetsen bevatten voor de meerderheid dezelfde items als de papieren toetsen van hetzelfde niveau.
- Eenzelfde opgave mag maximaal twee keer voorkomen in het digitale volgsysteem Spelling 3.0.
- Geen opgaven met DIF ten opzichte van de papieren toetsen opnemen.
  Er is bij een digitale opgave sprake van DIF wanneer het bij gefixeerde itemparameters in de papieren versie niet mogelijk is het digitale item op dezelfde schaal te kalibreren.
- De gemiddelde moeilijkheid van de digitale toetsen is zoveel mogelijk gelijk aan die van de papieren toetsen vanwege eenzelfde toetsbeleving voor de leerlingen.
- De items van de digitale toetsen verwijzen naar precies dezelfde categorieën als de papieren toetsen.
- Net als bij de papieren toetsen komen in principe alle opgaven met een acceptabele moeilijkheid (in klassieke termen een p-waarde tussen 0,40 en 0,90) die door de betere spellers significant vaker goed worden gemaakt dan door de minder goede spellers (rir vanaf 0,20) in aanmerking voor opname in de definitieve digitale toetsen Spelling.

Aan de voorwaarden waarnaar deze uitgangspunten verwijzen, hebben we in grote lijnen kunnen voldoen. De digitale toets Spelling niet-werkwoorden B8/M8 heeft grote overlap met de corresponderende papieren toets. De opgaven in de toets B8/M8 komen niet voor in andere toetsen Spelling 3.0. De aantallen opgaven per categorie van de digitale toets B8/M8 zijn vrijwel gelijk aan die van de papieren variant. We zien alleen een verschil bij categorie 27 (één opgave meer in de digitale toets) en categorie 41 (één opgave minder in de digitale toets). De precieze aantallen opgaven per categorie staan in tabel 3.1a.

Tabel 3.1a Spellingcategorieën in de digitale toets Spelling 3.0 groep 8 niet-werkwoorden (met de aantallen in de papieren toets tussen haakjes)

Spelling niet-werkwoorden		
Categorie	**Omschrijving**	**B8/M8**
20/21	woorden met open/gesloten lettergreep	5 (5)
26	woorden waarin /s/ geschreven wordt als c	4 (4)
27	woorden waarin /k/ geschreven wordt als c	5 (4)
28	woorden beginnend met ’s of eindigend op ’s	4 (4)
35	woorden met een trema	5 (5)
37	samenstelling met tussen -e(n)-	5 (5)
39/40	Franse en Engelse leenwoorden	5 (5)
41	woorden waarin /t/ geschreven wordt als th	3 (4)
43	woorden waarin /ks/ geschreven wordt als x	5 (5)
46	woorden op -iaal, -ieel, -ueel, -eaal	5 (5)
47	stoffelijke bijvoeglijke naamwoorden	4 (4)

Ook bij de digitale toets Spelling werkwoorden B8/M8 komen de aantallen opgaven per categorie vrijwel volledig overeen met die van de papieren variant, behalve bij categorie 1.c (één opgave minder in de digitale toets) en categorie 2.c (één opgave meer in de digitale toets). De precieze aantallen opgaven per categorie staan in tabel 3.1b.

Tabel 3.1b Spellingcategorieën in de digitale toets Spelling 3.0 groep 8 werkwoorden (met de aantallen in de papieren toets tussen haakjes)

Spelling werkwoorden			
	Categorie	**Omschrijving**	**B8/M8**
1. o.t.t.	1.b	wel of geen -t achter een stam op -d	6 (6)
	1.c	bij inversie pv-ond: wel of geen -t achter een stam op -d (vraag of gebiedende wijs)	5 (6)
	1.d	homofone gevallen	5 (5)
2. o.v.t.	2.b	verdubbeling d of t bij zwak ww met stam op -d of -t	6 (6)
	2.c	geen -t bij sterk ww dat in 2e en 3e persoon eindigt op -d	6 (5)
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	6 (6)
	3.b	homofone gevallen	6 (6)
	3.c	zwakke werkwoorden met stam op -d, -t, -v of -z	6 (6)
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	4 (4)

Net als de papieren toetsen bevatten de digitale toetsen 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 digitale toetsen zijn de figuren in bijlage 1: p50- en p80-kanspunten van de opgaven in de digitale toetsen voor groep 8 in relatie tot de gemiddelde vaardigheidsscore voor het afnamemoment. In deze figuren is de verdeling van de opgaven over de taken van de toetsen visueel weergegeven. De balkjes in de figuren geven het p50- (onderkant van het balkje) en p80-kanspunt (bovenkant van het balkje) van elke opgave aan. Het p50-punt geeft de vaardigheidsscore aan waarbij er sprake is van een kans van 50% om een opgave goed te beantwoorden. In deze figuren is zichtbaar dat de toetsen opgaven bevatten van uiteenlopende moeilijkheidsgraad.

Bij de digitale toets Spelling niet-werkwoorden 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) opgenomen. De meeste opgaven hebben een gemiddelde moeilijkheid. Ook zijn er naar verhouding veel opgaven relatief gemakkelijk, zodat de leerlingen een prettige toetservaring beleven. Ook de digitale toets Spelling werkwoorden heeft een spreiding van makkelijke opgaven, opgaven van gemiddelde moeilijkheid en moeilijke opgaven. Wel is het aandeel relatief moeilijke opgaven groter dan bij de digitale toets Spelling niet-werkwoorden. Ditzelfde beeld zagen we bij de papieren toetsen.

In de tabellen 3.2a en 3.2b zijn de ranges en de gemiddelden weergegeven voor de p-waarden en de r_{it}-waarden van de items van de digitale toetsen groep 8. Bij alle toetsen is te zien dat de p-waarden liggen tussen 0,36 en 0,93. Er is gestreefd naar p-waarden van de items tussen 0,40 en 0,90. Op een paar items na is dit gelukt. Er is gestreefd naar een goede spreiding van moeilijkheid over de items. De gemiddelde moeilijkheid van de items voor niet-werkwoorden en werkwoorden is 0,67 cq. 0,65.

Bij geen enkele opgave ligt de r_{it}-waarde onder 0,20. De gemiddelde r_{it}-waarde is voor de digitale toetsen 0,37 cq. 0,40. Door de Cotan wordt een r_{it}-waarde hoger dan 0,30 gekwalificeerd als goed. Met een

gemiddelde van 0,37 of hoger is de itemkwaliteit van de toetsen goed te noemen. Bijlage 2 bevat een volledig overzicht van de p-waarden en de r_{it}-waarden van de items van de digitale toetsen.

Voor de volledigheid hebben we ook de ranges en de gemiddelden weergegeven voor de p-waarden en de r_{it}-waarden van de items van de papieren toets groep 8 voor zowel niet-werkwoorden als werkwoorden (zie de tabellen 3.3a en 3.3b die overgenomen zijn uit de wetenschappelijke verantwoording van de papieren toetsen). Te zien is dat de digitale toetsen enigszins moeilijker zijn dan de papieren toetsen. Dit heeft als consequentie dat een leerling in een digitale toets iets minder opgaven goed hoeft te hebben dan in een papieren toets van hetzelfde niveau om eenzelfde vaardigheidsscore te behalen.

Tabel 3.2a Range en gemiddelde van p- en R_{it}-waarden voor de digitale toets groep 8 van Spelling 3.0 niet-werkwoorden

	p-waarden		Rit-waarden		N items
	Range	**Gemiddelde**	**Range**	**Gemiddelde**	
M8	0,36 – 0,93	0,67	0,25 – 0,55	0,40	50

Tabel 3.2b Range en gemiddelde van p- en R_{it}-waarden voor de digitale toets groep 8 van Spelling 3.0 werkwoorden

	p-waarden		Rit-waarden		N items
	Range	**Gemiddelde**	**Range**	**Gemiddelde**	
M8	0,39 – 0,84	0,65	0,24 – 0,58	0,37	50

Tabel 3.3a Range en gemiddelde van p- en R_{it}-waarden voor de papieren toets groep 8 van Spelling 3.0 niet-werkwoorden

	p-waarden		Rit-waarden		N items
	Range	**Gemiddelde**	**Range**	**Gemiddelde**	
M8	0,43 – 0,92	0,68	0,27 – 0,56	0,42	50

Tabel 3.3b Range en gemiddelde van p- en R_{it}-waarden voor de papieren toets groep 8 van Spelling 3.0 werkwoorden

	p-waarden		Rit-waarden		N items
	Range	**Gemiddelde**	**Range**	**Gemiddelde**	
M8	0,43 – 0,87	0,70	0,23 – 0,57	0,38	50

Hoewel de digitale toetsen dus niet identiek zijn aan de papieren toetsen, maakt het voor de resultaten niet uit of leerlingen de papieren of de digitale toetsen maken. De papieren en de digitale opgaven konden namelijk door middel van papier-digitaal vergelijkingsonderzoek in één opgavenbank ondergebracht worden; dat wil zeggen dat ze één en dezelfde vaardigheidsschaal representeren. Elke verzameling opgaven, of dit nu digitale opgaven zijn of opgaven op papier, is dan geschikt om die vaardigheid te toetsen, mits de betreffende verzameling min of meer is afgestemd op het niveau van de doelgroep. Zie voor een verantwoording hoofdstuk 4.

3.3 Statistische beschrijving

Voor het schatten van de vaardigheid van een leerling maken we bij de LVS-toetsen in het algemeen gebruik van twee berekeningswijzen. Bij de eerste berekeningswijze wordt uitgegaan van het aantal 'goed' op de toets en worden de opgaven níet gewogen met de discriminatie-index: elke opgave telt even zwaar mee in de berekening. Deze berekeningswijze maakt gebruik van de zogenoemde ongewogen score. Bij de tweede berekeningswijze worden de opgaven wél gewogen met hun discriminatie-index, er is dan sprake van gewogen scores. Statistisch gezien is de tweede berekeningswijze te prefereren: de gewogen score (bij gebruik van OPLM) is namelijk een voldoende statistiek voor de (latente) vaardigheid. Met andere woorden, alle informatie over de vaardigheid kunnen we bepalen met behulp van de gewogen score. Voor de schattingswijze van de (latente) vaardigheid waarbij gebruik gemaakt wordt van de ongewogen score geldt dat we (een klein beetje) informatie verliezen. Bovendien geldt dat de schattingen asymptotisch identiek zijn én dat beide schattingen gelijk zijn aan de ware vaardigheid. Het gebruik van de gewogen score heeft één voordeel boven het gebruik van de ongewogen score: de standaardfout van de schatting is aanzienlijk kleiner. Een nadeel is echter dat voor het berekenen van de gewogen score het gehele antwoordpatroon van de leerling nodig is. Dat vraagt voor scholen die de papieren toetsen handmatig verwerken een grote tijdsinspanning. Bij de digitale toetsen speelt dit nadeel niet, omdat de antwoorden op de computer automatisch worden opgeslagen en verwerkt. Daarom wordt bij het schatten van de vaardigheidsscores van digitale toetsen altijd gebruikgemaakt van de gewogen scores.

Voor de verschillende boekjes in het kalibratieonderzoek papier-digitaal zijn de correlaties tussen de schattingen met de beide genoemde berekeningswijzen alle groter dan .99. Er zijn voor deze situatie T-testen uitgevoerd: voor alle leerlingen geldt dat er geen significant verschil is tussen beide schattingen. Voor de beide reeksen (per afnamemoment) is ook gekeken naar het teken van de verschilscores: vaardigheidsscore gewogen papier – vaardigheidsscore gewogen digitaal én vaardigheidsscore gewogen digitaal – vaardigheidsscore digitaal ongewogen. In beide gevallen is het teken ongeveer even vaak positief als negatief. Voor de vergelijking van de gewogen versus de ongewogen vaardigheidsscores geldt bijvoorbeeld dat in ongeveer de helft van de gevallen de schatting op basis van de gewogen score kleiner is dan die op basis van de ongewogen scores. We kunnen hieruit concluderen dat het voor het schatten van de vaardigheid geen verschil maakt welke van de beide schattingsmethoden (gewogen of ongewogen) wordt gebruikt. Omdat het bij digitale toetsen makkelijk te implementeren is om de gewogen score te gebruiken en omdat deze theoretisch preciezer is, wordt bij digitale toetsen de voorkeur gegeven aan de gewogen score.

In de tabellen 3.4a en 3.4b is te zien dat de vaardigheidsverdelingen exact hetzelfde zijn als die van de verdelingen van de papieren versie. De reden hiervoor is dat voor de normering van de digitale toetsen de normen van de papieren toetsen zijn aangehouden. De gegevens zijn gebaseerd op 993 leerlingen voor M8 niet-werkwoorden en 890 leerlingen voor M8 werkwoorden. Dit betreft de aantallen leerlingen van de normering van de papieren toetsen. De waarden laten zien dat de vaardigheidsverdeling bij benadering normaal is. De figuren 3.1 en 3.2 met de verdeling van de vaardigheidsscores laten dit ook zien. Ook deze figuren zijn precies zo opgenomen in de wetenschappelijke verantwoording van de papieren toetsen Spelling 3.0 groep 8. Voor de duidelijkheid geven we deze figuren hier nogmaals weer.

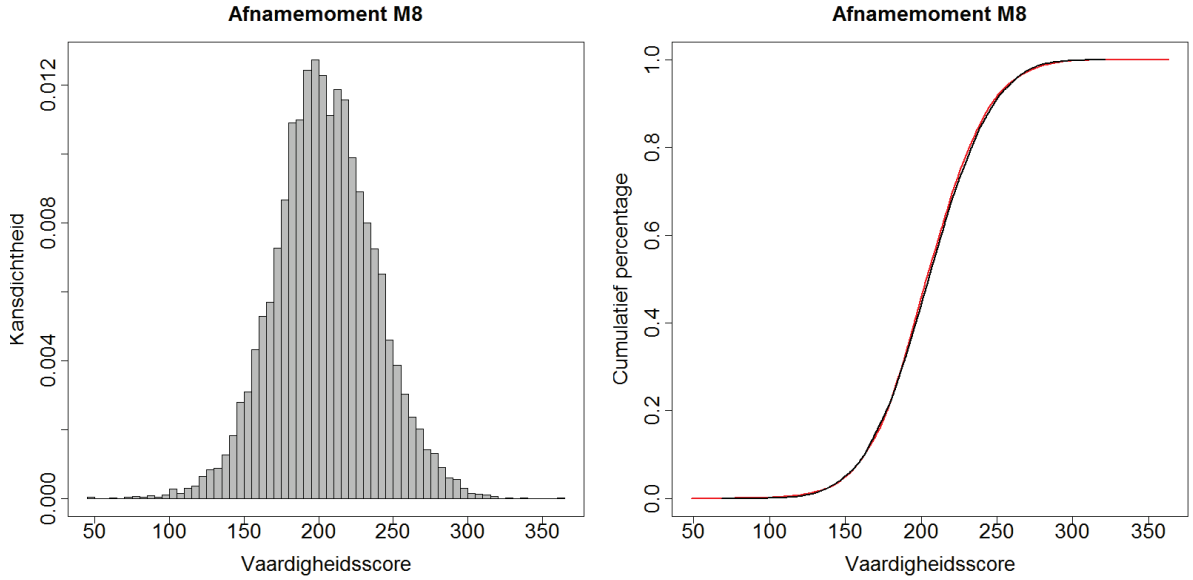
Tabel 3.4a Beschrijvende gegevens digitale toets Spelling niet-werkwoorden groep 8 op de gewogen scoreschaal en de vaardigheidsschaal

	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
M8 nww gewogen score	111,2	31,7	-0,28	-0,600
M8 nww vaardigheid	366,8	27,1	0,06	0,24

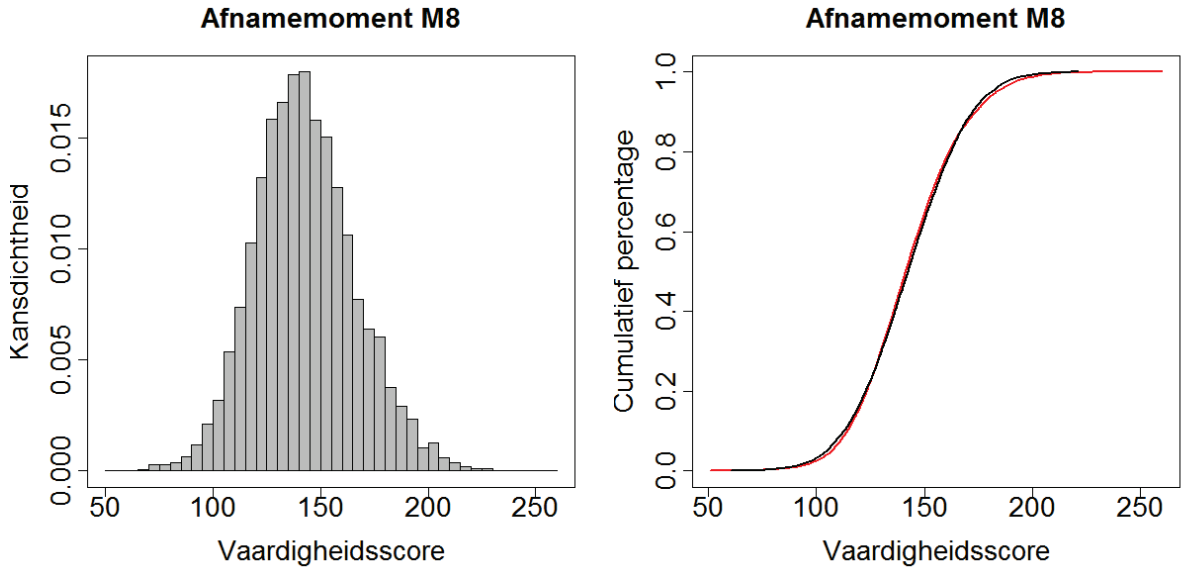
Tabel 3.4b Beschrijvende gegevens digitale toets Spelling werkwoorden groep 8 op de gewogen scoreschaal en op de vaardigheidsschaal

	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
M8 ww gewogen score	122,0	38,3	-0,63	-0,40
M8 ww vaardigheid	142,7	23,4	0,28	0,24

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



Figuur 3.2 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling M8 werkwoorden



4 Kalibratie en normering

4.1 Rationale van de kalibratieonderzoeken

Aan het begin van het toetsontwikkelingsproces van de LVS-toetsen Spelling 3.0 groep 8, zijn in 2012 en 2013 opgaven geconstrueerd. Deze opgaven zijn allereerst in *papieren* versie onderzocht in kalibratie-onderzoeken in 2015 en 2016 en – na opname van de opgaven in de uit te geven 'papieren' toetsen – in normeringsonderzoeken in 2017. Zie voor het kalibratie- en normeringsonderzoek van de papieren items voor groep 8 paragraaf 4.2 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019).

In 2018 hebben vervolgens kalibratieonderzoeken voor de *digitale* items plaatsgevonden. Dit gebeurde in de vorm van papier-digitaal vergelijkingsonderzoek. De hoofdvraag in deze kalibratieonderzoeken was: meten de papieren items en de digitale items dezelfde vaardigheid? We kunnen deze vraag bevestigend beantwoorden als de papieren en digitale items op dezelfde schaal blijken te passen.
Hieronder in paragraaf 4.2 beschrijven we deze papier-digitaal vergelijkingsonderzoeken en rapporteren we de resultaten. De conclusie is dat de digitale items dezelfde vaardigheid meten als de papieren items.
Dit betekent dat de papieren en de digitale items op één schaal passen, dat de papieren en digitale versies van toetsen van hetzelfde niveau leiden tot dezelfde vaardigheidsschatting en dat dezelfde normering gehanteerd kan worden. De normering lag al vast voor de papieren toetsen. We nemen hieronder in paragraaf 4.3 de normeringsgegevens in verkorte versie over van de verantwoording van de papieren toetsen, omdat er sprake is van één normering die zowel geldt voor de papieren toetsen als de digitale toetsen.

In dit hoofdstuk bespreken we besproken hoe de digitale en papieren opgaven op de vaardigheidsschaal spelling passen. De papieren en de digitale opgaven vormen daarmee ook één opgavenbank. Dat betekent onder andere dat de vaardigheid van een leerling met elke willekeurige selectie van opgaven, ook een selectie van *digitale* opgaven, uit deze bank gemeten kan worden. Hoe nauwkeurig dat gebeurt hangt uiteraard af van het aantal opgaven en van de psychometrische eigenschappen van de opgaven, onder andere de moeilijkheid van de gekozen opgaven in relatie tot de vaardigheid van de leerling. Daarmee is meteen ook het tweede doel van de hier gerapporteerde analyses gegeven: het onderbouwen van de keuze van de items die - gegeven de doelgroep - het beste in de digitale toetsen passen. We kunnen laten zien dat de digitale toetsen voor groep 8 uiteindelijk opgaven bevatten met psychometrisch goede eigenschappen (zie tabel 5.1). Deze psychometrische eigenschappen komen overeen met die van de papieren versies. We kunnen met de digitale opgaven dus even goed de vaardigheid spelling meten als met de papieren opgaven. Er is daarom geen onderzoek nodig naar de equivalentie van de scores op de verschillende versies: de scores zijn immers per definitie uitwisselbaar. Dit betekent ook dat de normen voor de papieren en de digitale toetsen hetzelfde kunnen en moeten zijn. Immers, met beide toetsen meten we dezelfde vaardigheid bij eenzelfde populatie.

4.2 Kalibratieonderzoek digitale items

4.2.1 Opzet van het kalibratieonderzoek voor de digitale items

In aparte kalibratieonderzoeken is onderzocht of de digitale items bij de papieren items op de schaal Spelling niet-werkwoorden dan wel werkwoorden passen. In januari 2018 vond het papier-digitaal-kalibratieonderzoek M8 niet-werkwoorden en M8 werkwoorden plaats.

Alle opgaven van de papieren toetsen 3.0 groep 8 zijn omgezet naar digitale versies, aangevuld met reserve-opgaven. In een papier-digitaal onderzoek is het design onvolledig: in tegenstelling tot een volledig

papieren onderzoek overlappen de boekjes slechts gedeeltelijk. De echte link om de boekjes op één schaal te krijgen, verloopt via de papieren uitgave. In de tabellen met afnamedesigns (tabel 4.1 en 4.2) zijn de boekjes van de papieren uitgave ook opgenomen. Hier is te zien dat de overlap via de papieren uitgave ervoor zorgt dat de afnamedesigns verbonden zijn. Alle nieuwe (digitale) taken zijn immers afgenomen bij leerlingen die ook de papieren Starttaak van LVS Spelling tweede generatie gemaakt hebben. Omdat de spellingvaardigheid van een leerling niet verandert tijdens een toets, kan de vaardigheid op het papieren gedeelte van de toets vergelijkbaar worden geacht met de vaardigheid op het digitale gedeelte van de toets. Hierdoor zijn de twee afnamemethoden verbonden. Voor de digitale items zijn wel eigen moeilijkheids- en discriminatieparameters geschat. Want ook al zijn de items van de papieren starttaak en de digitale starttaak inhoudelijk gelijk, door de verschillende afnamemethoden kunnen ze niet beschouwd worden als dezelfde items. Het is immers niet noodzakelijkerwijs zo dat de papieren versie en digitale versie van een item precies even moeilijk zijn en/of even goed discrimineren. We bespreken hieronder opzet en design voor de kalibratieonderzoeken.

Medio 8 niet-werkwoorden
In het papier-digitaal onderzoek voor het ‘medio’ (M) afnamemoment van januari 2018 zijn 100 items voorgelegd aan 378 leerlingen van groep 8. Elke leerling maakte één digitale taak met 35 nieuwe items uit LVS Spelling 3.0: ofwel taak 1 3.0 ofwel taak 2 3.0. De taken 1 en 2 bestonden elk uit 25 gedigitaliseerde items van de papieren toets M8 3.0, aangevuld met 10 reserve-items. Daarnaast maakte elke leerling de Starttaak M8 van LVS tweede generatie, bestaande uit 30 items. Ongeveer de helft van de leerlingen maakte de papieren Starttaak van de tweede generatie (de leerlingen die toegewezen waren aan boekje 1 en 2) en ongeveer de helft maakte de digitale Starttaak van de tweede generatie (de leerlingen die toegewezen waren aan boekje 3 en 4). Er is voor gezorgd dat geen enkele leerling beide versies (papier en digitaal) van eenzelfde item kreeg voorgelegd.

Doordat de helft van de onderzochte groep de Starttaak van de tweede generatie op papier maakte, konden we de papieren en digitale items vergelijken: passen ze op een en dezelfde schaal? Doordat ook een grote groep leerlingen beide taken digitaal maakte, waren we in staat de beste *digitale* items te selecteren voor de definitieve digitale toetsen.
De items waren verdeeld over vier verschillende opgavenboekjes in een onvolledig, maar ‘verbonden’ design. De data van het papier-digitaal onderzoek werden gekoppeld aan de data die verzameld werden in het normeringsonderzoek voor de papieren uitgave M8 uit 2017. Om de koppeling te realiseren werden de data die verzameld zijn in het papier-digitaal onderzoek toegevoegd aan de dataset van het genoemde normeringsonderzoek (die dient voor de schaling van de items in de itembank Spelling).

Tabel 4.1 Afnamedesign proefonderzoek papier-digitaal M8 niet-werkwoorden

	Taak				**Aantal leerlingen**
	Papier	**Digitaal**			
Boekje	**Starttaak M8 LVS 2e generatie**	**Starttaak M8 LVS 2e generatie**	**Taak 1 3.0**	**Taak 2 3.0**	
1					112
2					101
3					77
4					88
Aantal leerlingen per taak	213	165	189	189	378

Medio 8 werkwoorden
In het papier-digitaal onderzoek voor het 'medio' (M) afnamemoment van januari 2018 zijn 85 items voorgelegd aan 347 leerlingen van groep 8. Elke leerling maakte één digitale taak met 30 nieuwe items uit LVS Spelling 3.0: ofwel taak 1 3.0 ofwel taak 2 3.0. De taken 1 en 2 bestonden elk uit 25 gedigitaliseerde items van de papieren toets M8 3.0, aangevuld met 5 reserve-items. Daarnaast maakte elke leerling de Starttaak M8 werkwoorden van LVS tweede generatie, bestaande uit 25 items. Ongeveer de helft van de leerlingen maakte de papieren Starttaak van de tweede generatie (de leerlingen die toegewezen waren aan boekje 1 en 2) en ongeveer de helft maakte de digitale Starttaak van de tweede generatie (de leerlingen die toegewezen waren aan boekje 3 en 4). Er is voor gezorgd dat geen enkele leerling beide versies (papier en digitaal) van eenzelfde item kreeg voorgelegd.

Doordat de helft van de onderzoeksgroep de Starttaak van de tweede generatie op papier maakte, konden we de papieren en digitale items vergelijken: passen ze op een en dezelfde schaal? Doordat ook een grote groep leerlingen beide taken digitaal maakte, waren we in staat de beste *digitale* items te selecteren voor de definitieve digitale toetsen.
De items waren verdeeld over vier verschillende opgavenboekjes in een onvolledig, maar 'verbonden' design. De data van het papier-digitaal onderzoek werden gekoppeld aan de data die verzameld werden in het normeringsonderzoek voor de papieren uitgave M8 werkwoorden uit 2017. Om de koppeling te realiseren werden de data die verzameld zijn in het papier-digitaal onderzoek toegevoegd aan de dataset van het genoemde normeringsonderzoek (die dient voor de schaling van de items in de itembank Spelling).

Tabel 4.2 Afnamedesign proefonderzoek papier-digitaal M8 werkwoorden

	Taak				**Aantal leerlingen**
	Papier	**Digitaal**			
Boekje	**Starttaak M8 LVS 2e generatie**	**Starttaak M8 LVS 2e generatie**	**Taak 1 3.0**	**Taak 2 3.0**	
1					67
2					127
3					74
4					79
Aantal leerlingen per taak	194	153	141	206	347

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. Hoe succesvol deze operatie is, kan statistisch getoetst worden. Eenvoudig gezegd, we schatten in OPLM met de CML-methode de itemparameters en controleren 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 θ. Dit betekent dat alle informatie in de

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

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

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

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

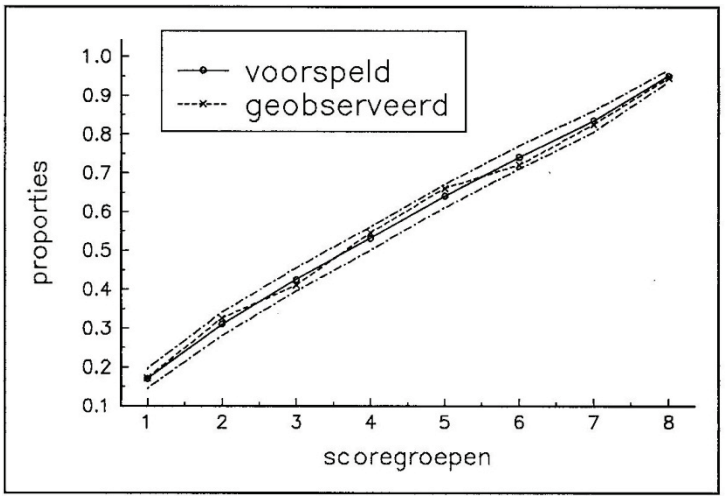
De stappen 1 tot en met 5 worden een aantal malen doorlopen tot het resultaat bevredigend is. Afhankelijk van de uitkomsten kunnen items worden verwijderd. Ook inhoudelijke overwegingen spelen een rol in dit beslissingsproces.

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 hiervoor 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 4.1 (zie Staphorsius, 1994, blz. 239). Figuur 4.1 beeldt voor een opgave de gegevens af waarop de betreffende S-toetsen gebaseerd zijn (zie handleiding OPLM: Verhelst; 1992). Ten behoeve van deze toetsing wordt de totale groep van leerlingen die een verzameling opgaven gemaakt heeft, ingedeeld in een aantal (meestal acht) score-groepen. Elke groep bestaat uit leerlingen met een ongeveer even hoge score. De geobserveerde proporties juiste antwoorden van deze groepen (telkens gesymboliseerd door een x) zijn door de middelste stippellijn verbonden. De volle lijn daarentegen verbindt de proporties die op grond van de parameter-schattingen voorspeld kunnen worden. De twee buitenste lijnen geven het 95%-betrouwbaarheidsinterval

aan. De breedte van dit interval is in belangrijke mate afhankelijk van het aantal leerlingen dat de opgave heeft beantwoord. Uit de figuur blijkt heel duidelijk dat de geobserveerde proporties, zoals bedoeld, binnen het 95%-betrouwbaarheidsinterval van de (geschatte) voorspelde proporties liggen, en dit komt in grote lijnen overeen met een niet-significante S-toetsingsgrootheid (Verhelst et al., 1994).

Figuur 4.1 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 digitale 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 4.2 illustreren dat voor de toetsen voor groep 6 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 4.1 overeenkomt. Dit is 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. Dat laatste wordt nog beter duidelijk voor de voorbeelden van S-toetsen van de ‘totale’ kalibraties van groep 6. Hierin wordt nog beter duidelijk hoe goed parameters van de items die uitsluitend op papier afgenomen zijn passen bij de parameters die digitaal zijn afgenomen. Er is nauwelijks tot geen sprake van differentieel itemfunctioneren (DIF). Met de ‘totale’ kalibraties wordt overigens de volledige opgavenbank van papieren en digitale items van groep 8 bedoeld. Meer over deze ‘totale’ kalibraties staat in paragraaf 4.2.4.

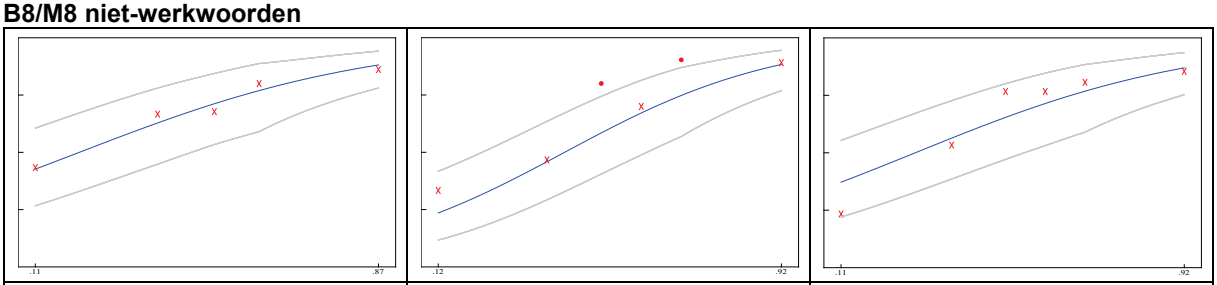
4.2.4 Totale kalibratie per groep

Om goed te kunnen bepalen of de digitale items (en hun parameters) passen bij de papieren parameters en om te bekijken of de items op één schaal passen is ook een ‘totale’ kalibratie uitgevoerd per groep, voor zowel Spelling niet-werkwoorden als voor Spelling werkwoorden. Dat wil zeggen dat voor alle items van de toets Spelling niet-werkwoorden van groep 8 een kalibratie werd uitgevoerd, alsook voor alle items van de toets Spelling werkwoorden van groep 8. Hierbij werden de parameters van de op papier afgenomen items

gefixeerd op de waarden zoals geschat in de normeringsonderzoeken van papier. Door vervolgens de parameters van de ‘digitale’ items op dezelfde kalibratieschaal te schatten als de ‘papieren’ items kan goed bepaald worden of deze items bij elkaar op een schaal passen. Ook kan bepaald worden of er sprake is van differentieel itemfunctioneren (DIF) van de digitale items ten opzichte van de papieren items. Er is bij een item sprake van DIF als de digitale versie ervan niet op dezelfde schaal kan worden gebracht. Hierna bespreken we daarom ook de passing van het meetmodel van de totale kalibratie per groep.
Omdat besloten is om geen aparte normeringen voor de digitale toetsen te ontwikkelen en de normering van de papieren items te gebruiken, is de passing van het meetmodel van de totale kalibratie per groep cruciaal om uitspraken te kunnen doen over de kwaliteit van de digitale toetsen LVS Spelling 3.0.

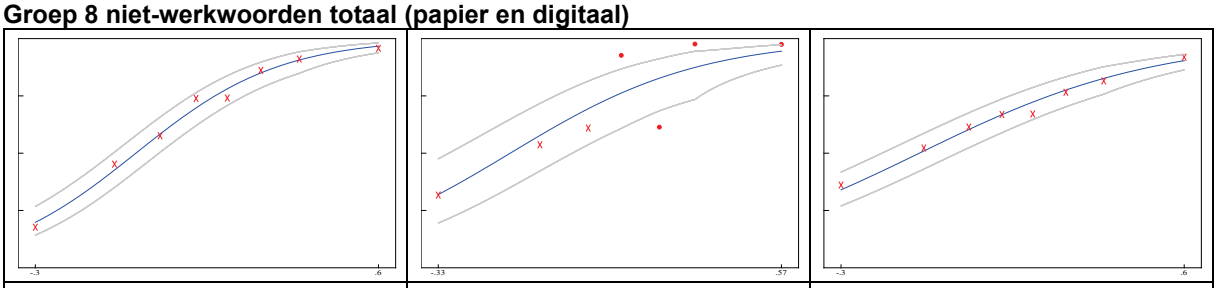
Figuur 4.2a Voorbeelden van S-toetsen voor de digitale toets Spelling 3.0 niet-werkwoorden groep 8 met de best passende, de slechtst passende en een qua passing representatieve opgave

B8/M8 niet-werkwoorden



Best passend	Slechtst passend	Representatieve passing

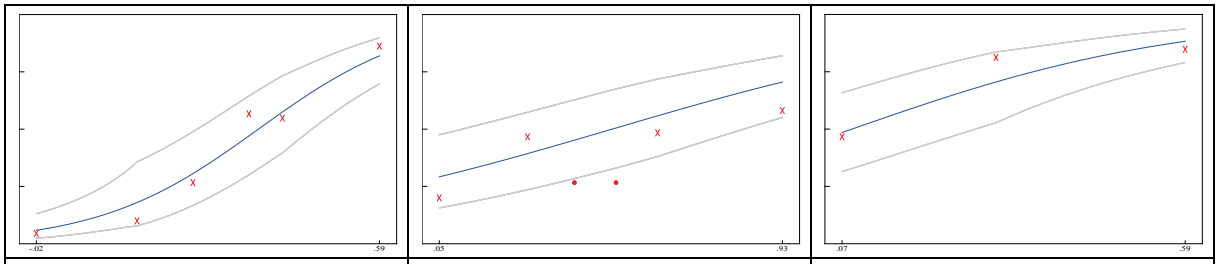
Groep 8 niet-werkwoorden totaal (papier en digitaal)



Best passend	Slechtst passend	Representatieve passing

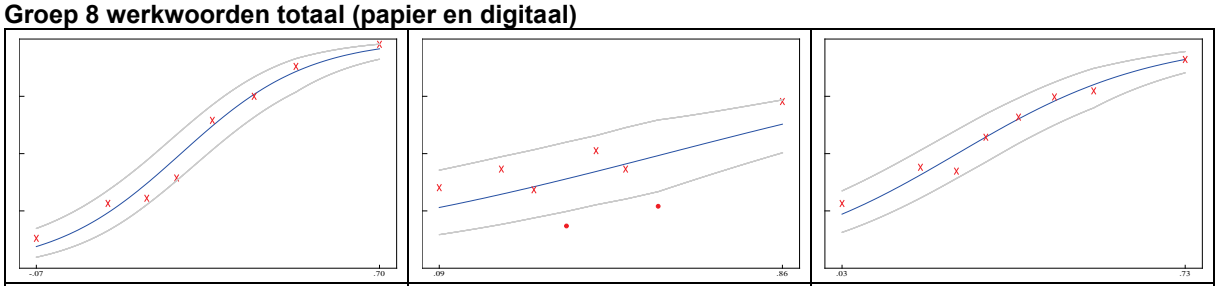
Figuur 4.2b Voorbeelden van S-toetsen voor de digitale toets Spelling 3.0 werkwoorden groep 8 met de best passende, de slechtst passende en een qua passing representatieve opgave

B8/M8 werkwoorden



Best passend	Slechtst passend	Representatieve passing

Groep 8 werkwoorden totaal (papier en digitaal)



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. Als we de S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn binnen het (0,1) interval, uiteraard met zo weinig mogelijk significante resultaten. Tabel 4.3a en 4.3b 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 digitale toetsen Spelling 3.0 groep 8. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan 0,01, respectievelijk 0,05. Het is duidelijk dat voor alle toetsen de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Deze resultaten geven 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 4.3a Verdeling van overschrijdingskansen bij S-toetsen voor digitale toetsen Spelling 3.0 groep 8 niet-werkwoorden en bij de totale kalibratie

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
B8/M8	0	2	2	3	1	4	5	7	7	3	10	6
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
Gr. 8 totaal	9	23	19	62	52	46	57	46	57	49	53	39

Tabel 4.3b Verdeling van overschrijdingskansen bij S-toetsen voor digitale toets Spelling 3.0 groep 8 werkwoorden en bij de totale kalibratie

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
B8/M8	2	5	2	2	5	4	5	1	3	5	1	5
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
Gr. 8 totaal	25	14	15	8	17	14	17	17	26	25	20	33

In tabel 4.4a en 4.4b zijn de R1c-waarden weergegeven voor dezelfde afnames waarvoor in tabel 4.3a en 4.3b 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 goede modelfit geldt als vuistregel dat R1c bij voorkeur niet groter zou moeten zijn dan ongeveer anderhalf maal het aantal vrijheidsgraden (df). Tot tweemaal het aantal vrijheidsgraden geldt het als een acceptabele passing. Uit tabel 4.4a en 4.4b blijkt dat de modelpassing voor de meeste toetsen goed is. Alle R1c-waarden zijn onder of rond anderhalf maal het aantal vrijheidsgraden. De significantie van de statistische toetsingen is bij de grote aantallen in de analyse nauwelijks informatief.

Tabel 4.4a R1c-waarden voor de digitale toets Spelling 3.0 B8/M8 niet-werkwoorden

Toetsversie	**R1c**	**df**	**p**
B8/M8	683,9	353	<0,01
Gr. 8 totaal	10428,4	6940	<0,01

Tabel 4.4b R1c-waarden voor de digitale toets Spelling 3.0 B8/M8 werkwoorden

Toetsversie	**R1c**	**df**	**p**
B8/M8	741,3	422	<0,01
Gr. 8 totaal	4229,1	2797	<0,01

Ten slotte bespreken we nog een methode om de modelpassing te verantwoorden die wordt besproken in het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer en Sijtsma, 2010, p. 40).
Het betreft hier een poging om de nauwkeurigheid van de itemparameterschattingen te beordelen op basis van een constante (in het COTAN Beoordelingssysteem met ‘c’ aangeduid) die weergeeft hoe de relatie is tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie. Het beoordelingssysteem geeft ook richtlijnen voor het beoordelen van de grootte van deze ‘c’. Deze dient te worden beoordeeld als ‘goed’ als de waarde lager is dan of gelijk aan 0,20. Waarden tussen 0,30 en 0,40 kunnen nog als voldoende worden beschouwd.
De waarden voor deze constante zijn weergegeven in tabel 4.5a en 4.5b. De gemiddelde waarden van de constante zijn goed te noemen. Zes items van B8/M8 werkwoorden zijn boven de 0,3 en onder de 0,4. Van B8/M8 niet-werkwoorden is één item boven de 0,3 en onder de 0,4.

De conclusie mag luiden dat we ook op basis van deze analyse de kalibratie geslaagd kunnen noemen.

Tabel 4.5a Nauwkeurigheid van de itemparameterschattingen (constante 'c') voor digitale toets Spelling 3.0 groep 8 niet-werkwoorden

Toetsversie	Constante 'c'	
	Range	Gemiddelde
B8/M8 nww	0,111 – 0,329	0,190

Tabel 4.5b Nauwkeurigheid van de itemparameterschattingen (constante 'c') voor digitale toets Spelling 3.0 groep 8 werkwoorden

Toetsversie	Constante 'c'	
	Range	Gemiddelde
B8/M8 ww	0,061 – 0,375	0,187

Ook op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de digitale toetsen Spelling 3.0 groep 8, voor zowel niet-werkwoorden als werkwoorden, de kalibratie geslaagd is. Belangrijker nog is de conclusie dat de kalibratie van de schaal waarop de papieren en de digitale items samen gekalibreerd zijn geslaagd is. Hieruit blijkt dat de papieren en de digitale items goed op een schaal passen en er geen sprake is van betekenisvol differentieel itemfunctioneren voor de digitale items in relatie tot de papieren items. Juist deze geslaagde kalibratie maakt het mogelijk om voor de digitale toetsen uit te gaan van de normen van de papieren toetsen.

4.3 De normering

De normering die wordt gebruikt voor de digitale toetsen Spelling is gelijk aan de normering van de papieren toetsen Spelling. Dit is mogelijk gezien de koppeling van het papier-digitaal kalibratieonderzoek aan de normeringsonderzoeken voor de papieren uitgaven via het in voorgaande paragraaf besproken design en de geslaagde kalibratie. De (papieren) normering is gebaseerd op de onderliggende (latente) verdeling van de vaardigheid op afnametijdstip M8. Bij de kalibratie is gebleken dat de 'digitale opgaven' op dezelfde schaal geplaatst konden worden als de 'papieren opgaven'. Daardoor kunnen we de eerder gevonden verdelingen van de vaardigheid van de normgroepen op deze schaal gebruiken. Voor het beschrijven van de normpopulatie kunnen we daarom gebruikmaken van de eerder gerapporteerde resultaten voor de papieren toetsversies.

De Expertgroep Toetsen PO had als oordeel dat voor de normering van Spelling 3.0 groep 8 op afname-moment M8 een representatieve steekproef is gebruikt. Ook zijn de gebruikte normgroepen groot genoeg en representatief voor de doelpopulatie, zowel op schoolniveau als leerlingniveau.
Deze normeringen worden in deze paragraaf in verkorte versie besproken. Voor een uitgebreide versie wordt verwezen naar paragraaf 4.3 van de wetenschappelijke verantwoording van de papieren toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019).

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

4.3.1 Opzet

Tijdens het embedded field normeringsonderzoek (zoals omschreven in paragraaf 4.2.1 van de Wetenschappelijke verantwoording van de papieren toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019)) werden data verzameld. Om deelnemers te werven voor het normerings-onderzoek 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.

Voor de normeringsonderzoeken M8 niet-werkwoorden en M8 werkwoorden zijn in eerste instantie 79 herhalingscholen (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 ongeveer 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. In tabel 4.6a en 4.6b zijn de uiteindelijke aantallen scholen en leerlingen in het (‘papieren’) normeringsonderzoek samengevat.

Tabel 4.6a 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 is lager dan dat van de oorspronkelijke normeringssteekproef doordat sommige scholen in de toepassing van het beschreven algoritme niet zijn geselecteerd.

Tabel 4.6b Aantal leerlingen die meegenomen zijn in de normering Spelling werkwoorden

Afnamemoment	Aantal leerlingen			Aantal scholen
	Steekproef	Dataretour 2e generatie	Normering	Normering
M8	654*	906	1560	87

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

N.B. De leerlingen die in het kalibratieonderzoek papier-digitaal zijn betrokken, maken geen deel uit van de normeringspopulatie.

4.3.2 Representativiteit

Door de werkwijze die werd gevolgd bij de normering is representativiteit van de normeringssteekproeven 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.
De steekproef is geanalyseerd in relatie tot de variabelen regio, urbanisatiegraad, schooltype en geslacht. De conclusie is dat de normeringssteekproeven een goede afspiegeling vormen van de populatie. Zie ook Tomesen, Wouda, Krämer & Horsels (2019).

4.3.3 Normeringsresultaten

Na de hierboven beschreven procedure doorlopen te hebben en de normeringssteekproef te hebben samengesteld, kon de normering worden bepaald. Naast het gemiddelde werden de percentielen bepaald. Dat gebeurde op basis van de verdeling van scores die werden gevonden in de normeringssteekproef zoals die is samengesteld op basis van het embedded field normeringsonderzoek en Cito-dataretour. Om de scores van leerlingen te kunnen vergelijken over de tijd worden vaardigheidsscores gebruikt. Uit de ruwe scores van de leerlingen uit het embedded field normeringsonderzoek en Cito-dataretour werden “plausible values” gegenereerd op de nieuw ontwikkelde vaardigheidsschaal. Deze “plausible values” representeren de volledige range aan vaardigheidsscores die een leerling zou kunnen hebben, gegeven de scores.
De “plausible values” geven niet alleen informatie over de geschatte vaardigheid maar ook over de onzekerheid die bij die schatting hoort (Keuning et al., 2014). De normering werd vervolgens gebaseerd op de “plausible values” van de leerlingen in de normeringssteekproef. Tabellen 4.7a en 4.7b geven de normgegevens voor de toetsen Spelling 3.0 groep 8.

Tabel 4.7a Normtabel op leerlingniveau voor Spelling 3.0 groep 8 niet-werkwoorden

Afname-moment	**M**	**SD**	**K**	**S**	**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 4.7b Normtabel op leerlingniveau voor Spelling 3.0 groep 8 werkwoorden

Afname-moment	**M**	**SD**	**K**	**S**	**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

De betreffende normeringstabellen zijn niet alleen van toepassing op de papieren versie van de toetsen Spelling 3.0 voor groep 8, maar ook op de hier verantwoorde digitale versie van deze toetsen.

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Betrouwbaarheid

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

In OPLAT wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ. Deze verwachte waarde wordt aangeduid met τ(θ). Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie worden aangegeven met het symbool Var(τ). Tussen θ en τ(θ) bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de toetsscore τ(θ) moet behalen (dat is alleen zo als de toets oneindig lang wordt). De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor met een eenmalige toetsafname niet meer zonder fout de waarde van θ bepaald kan worden. De variantie van de geobserveerde toetsscore wordt aangegeven met Var(t|τ(θ)), en door weer gebruik te maken van de distributie van θ in de populatie kan ook de gemiddelde variantie van de geobserveerde toetsscores berekend 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 vaardigheidsschaal Spelling niet-werkwoorden cq. werkwoorden, voor de digitale toetsen voor groep 8. In de eerste kolom staat de aanduiding van het afnamemoment. De tweede kolom geeft het aantal items van de toets weer en in de derde kolom staat de maximumscore die gehaald kan worden op de toets. Bij de papieren toetsen is de maximumscore gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. Bij de digitale toetsen gebruiken we echter de **gewogen** scores, zoals eerder al toegelicht. De vierde kolom geeft de geschatte gemiddelde score van de leerlingen op de toets. De vijfde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van de toets. De zesde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de toets is.

De betrouwbaarheidscoëfficiënten zijn hoog. Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen Spelling 3.0 groep 8) geeft de COTAN aan dat een betrouwbaarheidscoëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 voldoende, en een betrouwbaarheidscoëfficiënt hoger dan 0,80 goed (Evers, Lucassen, Meijer en Sijtsma, 2010). Op grond van dit criterium is de meetnauwkeurigheid van de toetsen (zeer) goed te noemen.

Tabel 5.1a Beschrijvende gegevens bij de digitale toets Spelling 3.0 groep 8 niet-werkwoorden

Toets	Aantal items	Maximum score	Gemiddelde	Standaard-meetfout	MAcc	Test-hertest (simulatie)
M8 nww	50	164	111,2	9,42	0,91	0,91

Tabel 5.1b Beschrijvende gegevens bij de digitale toets Spelling 3.0 groep 8 werkwoorden

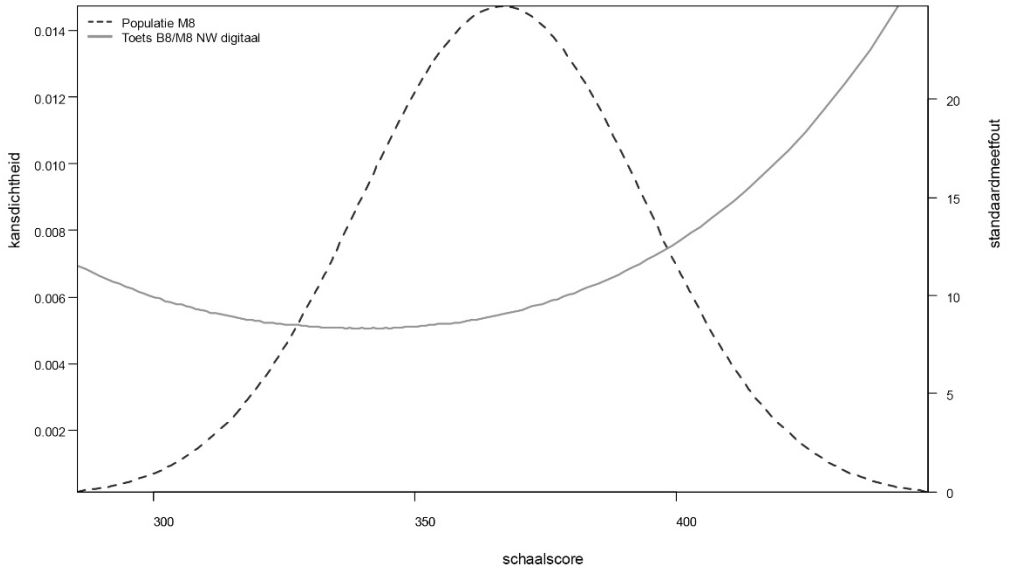
Toets	Aantal items	Maximum score	Gemiddelde	Standaard-meetfout	MAcc	Test-hertest (simulatie)
M8 ww	50	190	122,0	11,65	0,91	0,91

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

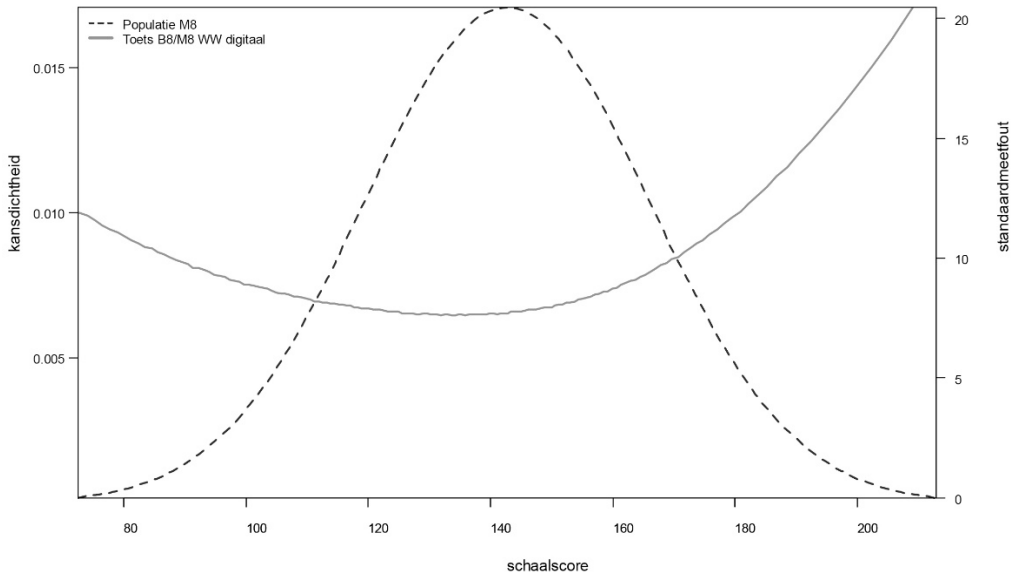
5.2 Nauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen en geven geen beeld van de lokale meetnauwkeurigheid van de digitale toetsen Spelling 3.0. De figuren 5.1a en 5.1b geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid van de toetsen. In deze figuren staat voor elke toets de grootte van de meetfout op de vaardigheidsschaal afgebeeld (met verdelingskenmerken zoals aangegeven in tabel 4.7a en 4.7b). Ook zijn de kansdichtheidfuncties voor de normgroep opgenomen. Deze laten zien hoe de vaardigheid van de leerlingen verdeeld is over de vaardigheidsschaal in de populatie die de toets gemaakt heeft. De figuren maken duidelijk dat de meetfout kleiner is in de lagere en gemiddelde vaardigheidsregionen dan in de hogere vaardigheidsregionen.

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



Figuur 5.1b Grootte van de meetfouten voor de digitale toets B8/M8 werkwoorden en de kansdichtheidsfunctie voor de M8-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 digitale toetsen B8/M8 niet-werkwoorden en werkwoorden zien hoe vaak de werkelijke vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat tabel 5.2 zien dat 83 procent van de leerlingen die halverwege groep 8 op basis van de digitale B8/M8-toets Spelling 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 aangemerkt is, met andere woorden, 83 procent. Verder laat de linkerkant van tabel 5.2 zien dat 16,6 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 digitale 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	83,0	16,6	0,4	0,0	0,0	E	79,0	20,6	0,4	0,0	0,0
IV	12,5	63,3	23,0	1,1	0,0	D	10,5	63,4	25,8	0,2	0,0
III	0,2	18,2	56,8	23,8	0,9	C	0,1	11,5	67,6	20,5	0,3
II	0,0	1,0	21,2	58,2	19,5	B	0,0	0,1	17,3	65,2	17,4
I	0,0	0,0	1,1	19,9	79,0	A	0,0	0,0	0,3	17,7	82,0

Tabel 5.3 Betrouwbaarheidstabel digitale 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	81,5	18,0	0,5	0,0	0,0	E	76,0	23,3	0,7	0,0	0,0
IV	13,8	62,0	23,1	1,1	0,0	D	11,9	61,1	26,8	0,3	0,0
III	0,3	19,5	56,7	22,8	0,7	C	0,2	13,1	67,4	19,2	0,2
II	0,0	1,0	21,2	58,9	18,9	B	0,0	0,1	17,6	65,2	17,0
I	0,0	0,0	0,9	18,9	80,2	A	0,0	0,0	0,2	16,3	83,4

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 (in Wheadon & Stockford, 2011). 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 de toetsen groep 8 zijn te vinden in tabel 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 de tabellen 5.4a en 5.4b aannemelijk dat de uitkomsten duidelijk in lijn zijn met het ambitieniveau zoals dat geformuleerd is door Pilliner (in Wheadon & Stockford, 2011) of zelfs boven dit ambitieniveau uitstijgen. Gemiddeld gezien scoort, afhankelijk van de toets en de gekozen indeling in scoregroepen, 99,0 tot 99,7 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 67,9 tot 71,4 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore gemiddeld gezien in zo'n 70 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. De resultaten stemmen hiermee tot grote tevredenheid: het percentage misclassificaties is zeer beperkt.
Op basis van bovenstaande gegevens concluderen we dat op basis van de digitale toetsen Spelling 3.0 groep 8 de leerlingen op een betrouwbare manier ingedeeld kunnen worden in normgroepen. Deze indeling voldoet over het algemeen uitstekend gegeven het doel van de toets. Uiteraard dienen de gebruikers rekening te houden met het gegeven dat er altijd sprake zal zijn van misclassificatie; veelal van maximaal één niveau verschil.

Tabel 5.4a Samenvattende indices digitale toets B8/M8 niet-werkwoorden

	Toets B8/M8, afnamemoment M8	
	Scoregroep I t/m V	**Scoregroep A t/m E**
Marginal classification accuracy	68,1	71,4
Accuracy plus/minus 1 niveau	99,0	99,7

Tabel 5.4b Samenvattende indices digitale toets B8/M8 werkwoorden

	Toets B8/M8, afnamemoment M8	
	Scoregroep I t/m V	**Scoregroep A t/m E**
Marginal classification accuracy	67,9	70,6
Accuracy plus/minus 1 niveau	99,1	99,7

6 Validiteit

Voor de verantwoording van de validiteit verwijzen we naar hoofdstuk 6 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019). Alles wat beschreven staat in dit hoofdstuk gaat ook op voor de digitale toetsen Spelling groep 8. Daarbij geldt dat ook de modelfit voor de digitale opgaven bevredigend is (zie paragraaf 4.2.3) en dat daarmee net als bij de papieren versie voldaan wordt aan eisen van unidimensionaliteit als waarborg voor de constructvaliditeit van de toetsen.

7 Samenvatting

In dit hoofdstuk vatten we samen wat in de voorafgaande hoofdstukken is besproken.
In hoofdstuk 1 is aangegeven dat het hier om een *aanvulling* gaat bij de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8 (Tomesen, Wouda, Krämer & Horsels, 2019). Deze aanvulling heeft uitsluitend betrekking op de *digitale* toetsen Spelling 3.0 voor groep 8, afnamemoment M8. Net als de papieren LVS-toetsen Spelling 3.0 groep 8 vormen de digitale LVS-toetsen Spelling 3.0 voor groep 8 een hulpmiddel om vast te stellen in hoeverre leerlingen kunnen spellen. De toetsen kunnen, in samenhang met de (papieren en digitale) toetsen Spelling 3.0 voor de andere leerjaren, worden gebruikt om de spellingvaardigheid van leerlingen in het primair en speciaal onderwijs in kaart te brengen en om hun ontwikkeling te volgen.

Voor de inhoudelijke aspecten verwijzen we naar de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8. Meetpretentie, gebruiksdoel en functie zijn identiek voor de papieren en digitale toetsen. Specifieke uitgangspunten bij het samenstellen van de digitale toetsen zijn in hoofdstuk 3 beschreven. Dit hoofdstuk bevat ook een beschrijving van enkele psychometrische kenmerken.

In hoofdstuk 4 verantwoorden we de kalibratie- en normeringsonderzoeken. De kalibratieonderzoeken zijn uitgevoerd in de vorm van papier-digitaal vergelijkingsonderzoeken. Hieruit blijkt dat de digitale items op dezelfde schaal passen en daardoor dezelfde vaardigheid meten als de papieren items. De modelfit voor de digitale items is bevredigend en daarmee is voldaan aan de eisen van unidimensionaliteit.
De normering lag al vast voor de papieren items; die hebben we ook aangehouden voor de digitale items.

In hoofdstuk 5 is over de betrouwbaarheidscoëfficiënten gerapporteerd. Net als bij de papieren toetsen, zijn de betrouwbaarheidscoëfficiënten (MAcc's en testhertest) voor de digitale versie van de toetsen voor groep 8 zeer hoog: 0,91 voor beide toetsen. Verder zijn in dit hoofdstuk betrouwbaarheidstabellen opgenomen die de betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden, laten zien. Daarnaast geven we inzicht in de lokale betrouwbaarheid: de meetfout blijkt het kleinst te zijn in de lagere en gemiddelde vaardigheidsregionen.

Voor de verantwoording van de validiteit (hoofdstuk 6) is weer verwezen naar de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 groep 8.

8 Aanvullende literatuur

Cito (2018). *Cito Volgsysteem primair en speciaal onderwijs. Spelling 3.0 Groep 8*. Arnhem: Cito.

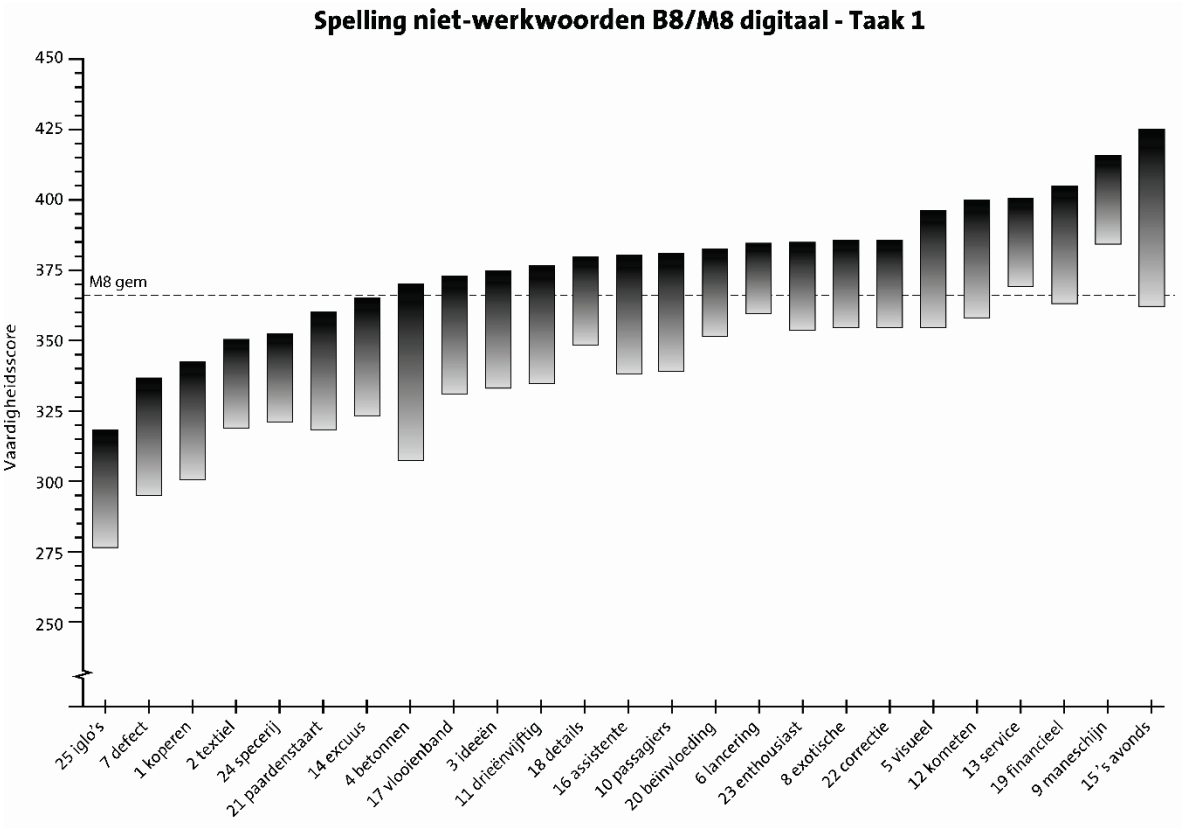
Cito (2019). *Spelling 3.0 Handleiding digitale toetsen*. Arnhem: Cito.

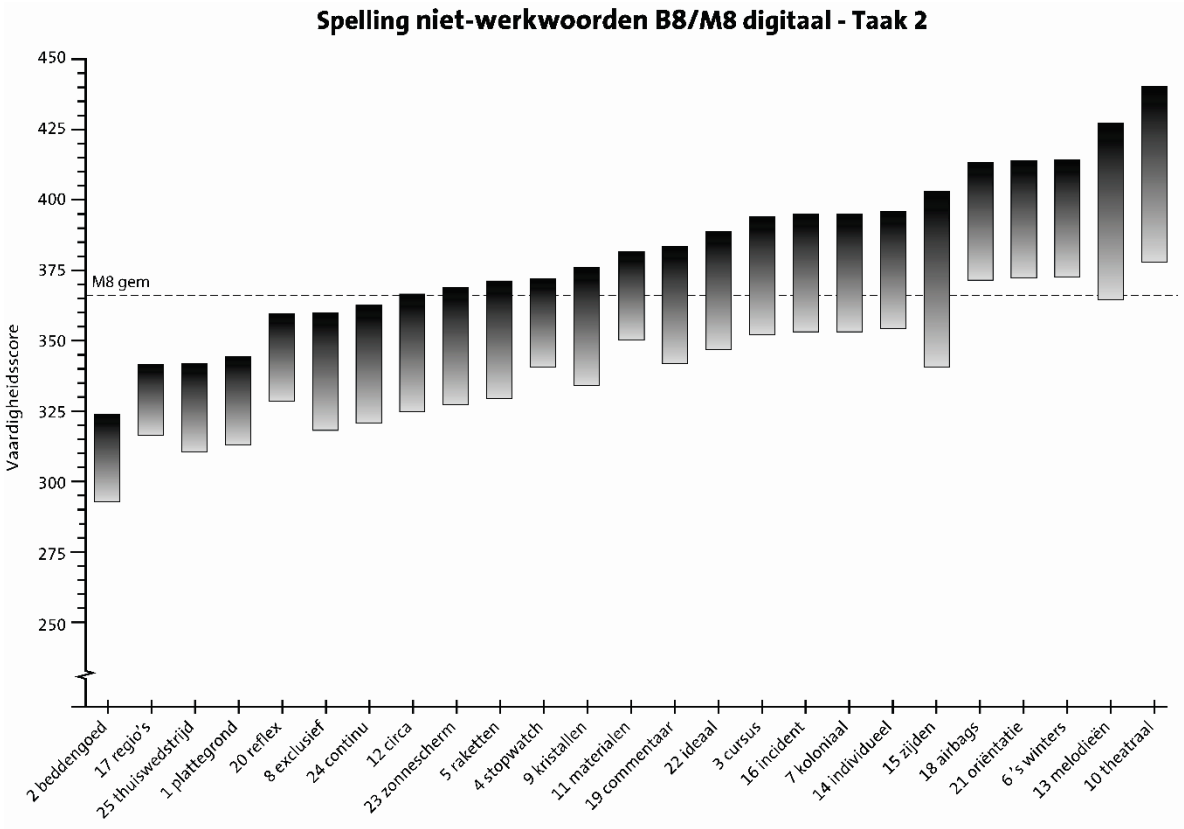
Tomesen, M., Wouda, J., Krämer, I. & Horsels, L. (2019). *Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 8*. Arnhem: Cito.

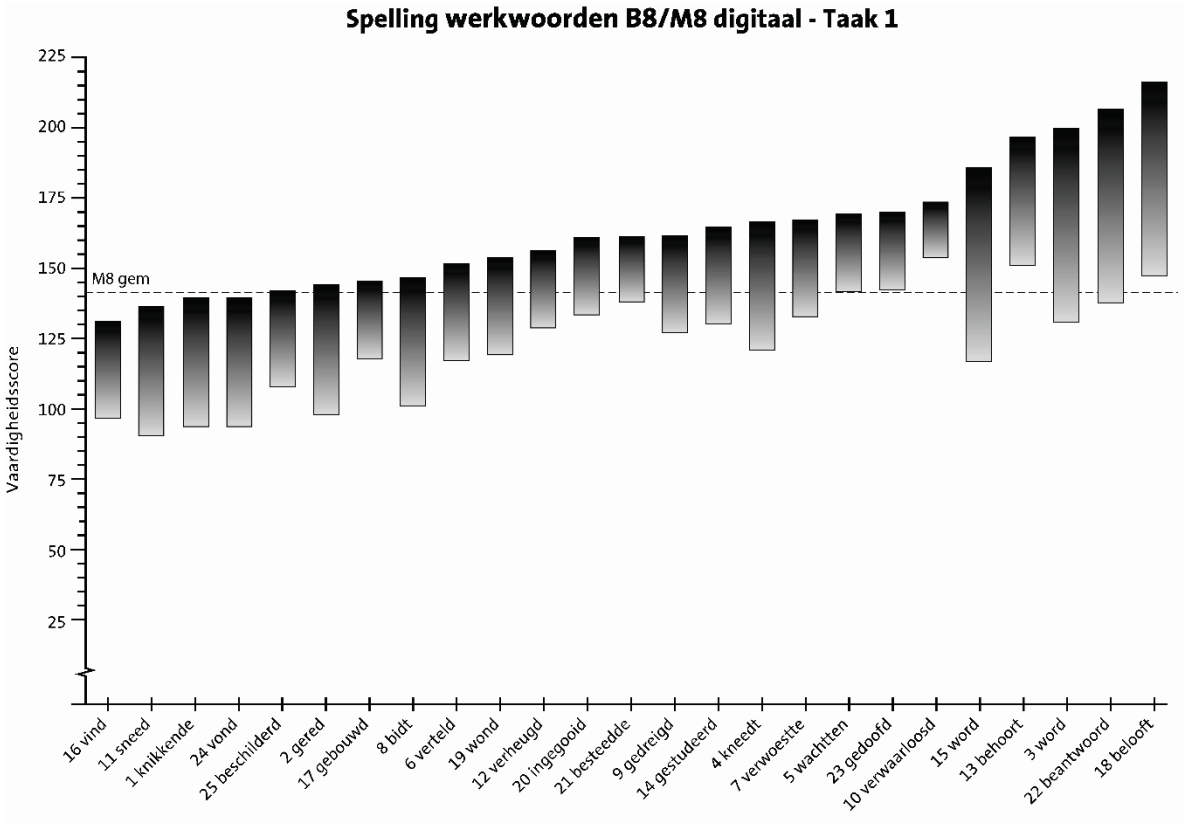
Wheadon, C. & Stockford, I. (2011*). Classification accuracy and consistency in GCSE and a level examinations offered by the assessment and qualifications alliance (AQA) november 2008 to june 2009*. Belfast: Office of Qualifications and Examinations Regulation.

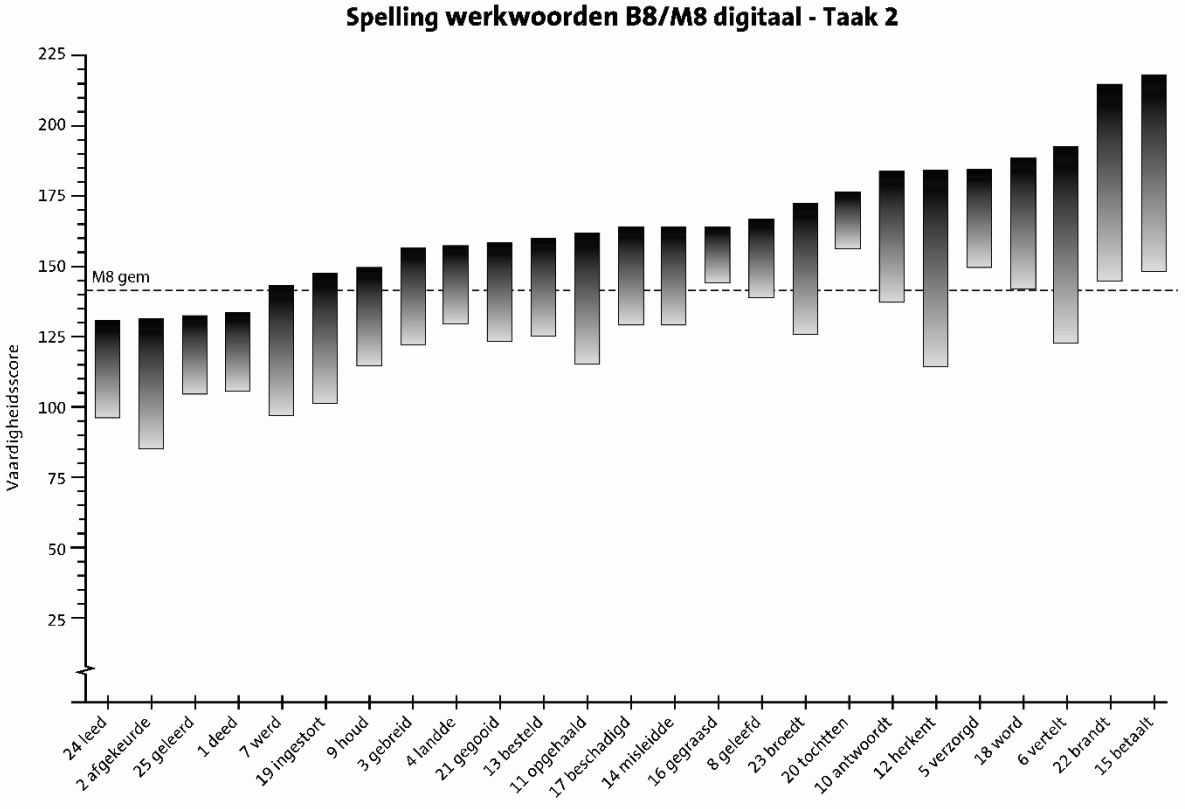
Bijlagen

Bijlage 1 Moeilijkheid van opgaven per taak in Spelling 3.0 digitaal groep 8






Spelling werkwoorden B8/M8 digitaal - Taak 1
Vaardigheidsscore
225
200
175
150
125
100
75
50
25
M8 gem
16 vind
11 sneed
1 knikkende
24 vond
25 beschilderd
2 gered
17 gebouwd
8 bidt
6 verteld
19 wond
12 verheugd
20 ingegooid
21 besteedde
9 gedreigd
14 gestudeerd
4 kneedt
7 verwoestte
5 wachtten
23 gedoofd
10 verwaarloosd
15 word
13 behoort
3 word
22 beantwoord
18 belooft


Spelling werkwoorden B8/M8 digitaal - Taak 2
Vaardigheidsscore
225
200
175
150
125
100
75
50
25
M8 gem
24 leed
2 afgekeurde
25 geleerd
1 deed
7 werd
19 ingestort
9 houd
3 gebreid
4 landde
21 gegooid
13 besteld
11 opgehaald
17 beschadigd
14 misleidde
16 gegraasd
8 geleefd
23 broedt
20 tochtten
10 antwoordt
12 herkent
5 verzorgd
18 word
6 vertelt
22 brandt
15 betaalt

Bijlage 2 Klassieke en IRT-indices van de opgaven in digitale toetsen Spelling 3.0 groep 8

Toets B8/M8 niet-werkwoorden: Normeringsmoment M8

Volgnr	P-Val	RIT	Beta	Info
1	0,865	0,316	-0,132	0,937
2	0,833	0,416	0,070	1,789
3	0,716	0,387	0,226	1,547
4	0,768	0,267	-0,059	0,661
5	0,583	0,406	0,461	1,813
6	0,566	0,550	0,517	4,128
7	0,883	0,300	-0,194	0,836
8	0,599	0,488	0,461	2,877
9	0,357	0,453	0,793	2,769
10	0,680	0,395	0,293	1,642
11	0,706	0,390	0,245	1,575
12	0,557	0,407	0,503	1,837
13	0,478	0,481	0,625	2,973
14	0,768	0,370	0,119	1,374
15	0,521	0,300	0,549	0,906
16	0,685	0,394	0,283	1,630
17	0,727	0,384	0,205	1,515
18	0,644	0,484	0,397	2,764
19	0,523	0,406	0,559	1,855
20	0,623	0,486	0,428	2,824
21	0,793	0,359	0,062	1,276
22	0,599	0,488	0,461	2,877
23	0,604	0,487	0,454	2,867
24	0,823	0,423	0,093	1,867
25	0,930	0,248	-0,398	0,546
26	0,860	0,395	0,008	1,579
27	0,929	0,312	-0,219	0,912
28	0,596	0,406	0,440	1,798
29	0,702	0,473	0,310	2,555
30	0,734	0,382	0,190	1,491
31	0,456	0,400	0,667	1,846
32	0,588	0,406	0,453	1,807
33	0,793	0,359	0,062	1,276
34	0,708	0,389	0,241	1,570
35	0,442	0,295	0,724	0,896
36	0,629	0,486	0,419	2,806
37	0,760	0,373	0,136	1,402
38	0,508	0,300	0,578	0,907
39	0,583	0,406	0,461	1,813
40	0,627	0,296	0,311	0,854
41	0,589	0,406	0,451	1,806
42	0,873	0,441	0,046	2,100
43	0,464	0,401	0,655	1,850
44	0,662	0,399	0,326	1,685
45	0,781	0,446	0,177	2,147
46	0,459	0,400	0,662	1,848
47	0,629	0,403	0,383	1,748
48	0,747	0,378	0,163	1,448
49	0,780	0,365	0,092	1,327
50	0,870	0,386	-0,020	1,489

Toets B8/M8 werkwoorden: Normeringsmoment M8

Volgnr	P-Val	RIT	Beta	Info
1	0,801	0,291	-0,062	1,306
2	0,781	0,299	-0,017	1,395
3	0,568	0,246	0,314	0,922
4	0,658	0,333	0,212	1,793
5	0,535	0,486	0,422	4,649
6	0,717	0,398	0,176	2,700
7	0,601	0,419	0,334	3,130
8	0,767	0,305	0,012	1,451
9	0,646	0,414	0,275	3,000
10	0,407	0,558	0,545	7,531
11	0,814	0,284	-0,093	1,243
12	0,651	0,477	0,294	4,303
13	0,463	0,340	0,517	1,964
14	0,621	0,417	0,308	3,077
15	0,631	0,241	0,174	0,877
16	0,837	0,342	-0,032	1,892
17	0,741	0,453	0,183	3,723
18	0,491	0,247	0,480	0,938
19	0,703	0,402	0,196	2,767
20	0,610	0,483	0,341	4,476
21	0,573	0,537	0,388	6,073
22	0,536	0,247	0,384	0,934
23	0,528	0,486	0,429	4,656
24	0,801	0,291	-0,062	1,306
25	0,777	0,375	0,081	2,344
26	0,825	0,409	0,056	2,902
27	0,836	0,272	-0,149	1,132
28	0,688	0,405	0,218	2,838
29	0,651	0,477	0,294	4,303
30	0,474	0,419	0,492	3,238
31	0,608	0,244	0,226	0,897
32	0,787	0,297	-0,031	1,366
33	0,569	0,486	0,386	4,594
34	0,736	0,392	0,147	2,597
35	0,558	0,342	0,371	1,950
36	0,692	0,326	0,153	1,705
37	0,644	0,240	0,144	0,864
38	0,664	0,411	0,251	2,936
39	0,634	0,416	0,291	3,039
40	0,491	0,247	0,480	0,938
41	0,523	0,577	0,440	7,740
42	0,635	0,415	0,290	3,036
43	0,528	0,343	0,417	1,968
44	0,767	0,305	0,012	1,451
45	0,389	0,553	0,561	7,439
46	0,676	0,408	0,234	2,888
47	0,506	0,247	0,448	0,938
48	0,631	0,337	0,257	1,851
49	0,840	0,340	-0,038	1,866
50	0,831	0,405	0,045	2,830

Cito
Amsterdamseweg 13
6814 CM Arnhem
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11

Fotografie: Gijs Versteeg