Wetenschappelijke verantwoording Spelling 3.0 digitaal voor groep 3

Aanvulling bij de wetenschappelijke verantwoording van LVS-Spelling 3.0 voor groep 3

Marieke Tomesen, Jasper Wouda en Linda Horsels

Wetenschappelijke verantwoording Spelling 3.0 digitaal voor groep 3

Aanvulling bij de wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3

Marieke Tomesen
Jasper Wouda
Linda Horsels

Cito
Arnhem, 2017

Inhoud

1 Inleiding

Deze *Aanvulling bij de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3* heeft uitsluitend betrekking op de *digitale* toetsen Spelling 3.0 voor groep 3.
De inhoud van deze digitale toetsen komt grotendeels overeen met de inhoud van de papieren toetsen voor groep 3. Vandaar dat we voor de inhoudelijke aspecten grotendeels verwijzen naar de oorspronkelijke Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3 (Tomesen, Wouda, Mols & Horsels, 2015). 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 3 (Tomesen, Wouda, Mols & Horsels, 2015) en de inhoud van het (digitale) toetspakket Spelling groep 3 (Cito, 2014) 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 3. Het genoemde materiaal maakt een beoordeling van de digitale toetsen Spelling groep 3 mogelijk op de volgende aspecten:

- Uitgangspunten van de toetsconstructie
- De kwaliteit van het toetsmateriaal
- De kwaliteit van de handleiding
- Normen
- Betrouwbaarheid
- Validiteit

Het laatstgenoemde aspect betreft alleen begripsvaliditeit en 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 3 (Tomesen, Wouda, Mols & Horsels, 2015). De kwaliteit van het toetsmateriaal en van de handleiding is te bepalen door kennis te nemen van de inhoud van de (digitale) toetspakketten.

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 3 (Tomesen, Wouda, Mols & Horsels, 2015). Alles wat in hoofdstuk 2 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3 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 3 van het basisonderwijs en voor leerlingen in het speciaal basisonderwijs en in het speciaal onderwijs cluster 2 en 4;
- voor zowel 'midden leerjaar' (half januari/half februari) als voor 'einde leerjaar' (juni) zijn populatie-parameters bepaald;
- ze kunnen ook gebruikt worden voor leerlingen in andere leerjaren die werken op het niveau van groep 3;
- voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften is ook een extra toets beschikbaar: M3E3, die net wat moeilijker is dan de toets M3 en wat gemakkelijker dan de toets E3;
- 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 zijn ook digitale varianten beschikbaar. Dit betekent dat er voor groep 3 een digitale toets M3, een digitale toets M3E3 en een digitale toets E3 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/zij 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.

Wat betreft de doelgroep is er wel een verschil. Dat lichten we hieronder kort toe.

Doelgroep
De toetsen worden digitaal afgenomen in plaats van op papier. Om dat goed te kunnen doen, moeten de leerlingen bekend zijn met de letters op het toetsenbord. Het is daarbij niet zozeer van belang dat ze de letters snel weten te vinden (er is geen tijdsdruk bij de toetsafname, dus een leerling kan rustig even zoeken naar een bepaalde letter). Maar het is wel belangrijk dat ze weten dat op het toetsenbord hoofdletters staan in plaats van kleine letters (ze moeten dus weten dat de I de hoofdletter i is en niet de kleine l (die staat onder hoofdletter L)). Scholen worden op deze specificatie van de doelgroep gewezen in de handleiding in paragraaf 2.2 'Keuze 1: papier of digitaal?'.

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 3 (Tomesen, Wouda, Mols & Horsels, 2015) – geldt zowel voor de papieren toetsen als de digitale toetsen.

3 Beschrijving van de toetsen

3.1 Opbouw en structuur van de toetsen

De digitale toetsen Spelling 3.0 voor jaargroep 3 kennen – net als de papieren toetsen voor groep 3 – drie versies: M3, M3E3 en E3. Voor de toets M3 zijn de populatieparameters bepaald op 'midden leerjaar' groep 3 (M3). Voor de toetsen M3E3 en E3 zijn de populatieparameters bepaald op 'einde leerjaar' (E3). De toets M3E3 is een tussentoets, bedoeld voor leerlingen die zich minder snel ontwikkelen.

De digitale varianten van de toetsen bevatten net als de papieren varianten twee taken van 20 opgaven en zijn op dezelfde manier opgebouwd. De opgaventypen (woorddictee voor M3 en zinsdictee voor de andere toetsen) zijn 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 3 (Tomesen, Wouda, Mols & Horsels, 2015) 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.

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 gemaakt. Afhankelijk van het aantal beschikbare computers kunnen meerdere leerlingen gelijktijdig aan dezelfde toets werken. De leerling krijgt voorafgaand aan elke taak een uitleg over het maken van de digitale opgaven en maakt enkele oefenopgaven. Op die manier raakt de leerling vertrouwd met het type opgave, maar ook met de stem die de opgaven voorleest. Dat kan prettig zijn voor leerlingen die dialect spreken.
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 digitale toets M3 ziet de leerling op het scherm een plaatje met daarnaast een streep. Bij de digitale toetsen M3E3 en E3 ziet de leerling op het scherm alleen een streep. De computer leest de opgave voor. Daarna verschijnt de cursor op de streep, zodat de leerling het dicteewoord kan intikken. Het is niet mogelijk om tijdens het afspelen van de audio al een antwoord in te tikken. De leerling kan, indien gewenst, de opgave nog een keer afspelen door op de knop met het oortje te klikken. Ook kan hij het volume aanpassen met de knop naast het oortje. De leerling moet minimaal twee letters intypen voordat hij op de knop 'verder' kan klikken. Daarna kan hij de volgende opgave maken.
Op die manier maakt hij één voor één de opgaven; kinderen kunnen geen opgaven overslaan in de digitale toetsen.

In de toetsmap Spelling 3.0 groep 3 (Cito, 2014) 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, 2016), die voor scholen via Cito Portal gedownload kan worden.
Bij de digitale versies van de toetsen worden de antwoorden van de leerlingen door de computer gescoord en hoeft de leerkracht de toetsen dus 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 computers) als van meer inhoudelijke aard. Vooral voor leerlingen met concentratieproblemen, leerlingen die langzamer dan gemiddeld werken en leerlingen die afwezig waren bij de klassikale afname, kan een individuele, digitale afname prettig zijn. De leerling moet wel op een toetsenbord kunnen werken 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 allerlei rapportages, zoals leerlingrapporten en groeps-overzichten, en een foutenanalyse worden opgevraagd. Dit is ook mogelijk bij de papieren toetsen, maar dan moet de leekracht of IB'er eerst nog de toetsresultaten zelf invoeren in het Computerprogramma LOVS.

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 3 (Tomesen, Wouda, Mols & Horsels, 2015). Het voor groep 3 afzonderlijk uitgewerkte overzicht van spellingcategorieën (zie tabel 3.1) 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 door een professionele 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 zij aan de hand van voorbeeldaudio het spreektempo; dit ligt namelijk wat lager bij lagere groepen dan bij hogere groepen. Bij twijfel over de uitspraak van een dicteewoord werd het uitspraakwoordenboek geraadpleegd (zie www.woorden.org).

Samenstelling definitieve toetsen
In januari en juni 2014 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 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 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.

We hebben gekozen voor een onderzoeksdesign waarbij zo min mogelijk leerlingen nodig waren, maar dat tegelijkertijd ook restricties oplegde aan het aantal items dat we konden inzetten in het onderzoek (in hoofdstuk 4 zullen we het design en de implicaties daarvan nader toelichten). Bij het samenstellen van de

digitale toetsen hadden we daarom minder keus in items dan bij de papieren toetsen en waren we genoodzaakt om items dubbel op te nemen, dat wil zeggen in verschillende toetsen.

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.
- De reguliere digitale toetsen (M3 en E3) bevatten voor de meerderheid dezelfde items als de reguliere papieren toetsen van hetzelfde niveau.
- De digitale tussentoets M3E3 mag deels overlappen met de toets E3 en met items van de tweede generatie digitale spellingtoetsen.
- Een opgave mag maximaal twee keer voorkomen in het digitale volgsysteem Spelling 3.0. De opgave wordt dan bij voorkeur opgenomen in een tussentoets en een toets van het opvolgende of voorafgaande niveau.
- Een opgave mag bij voorkeur niet meer dan een half toetsniveau hoger of lager opschuiven t.o.v. de papieren uitgave. Bijvoorbeeld: een opgave die in de uitgave van de papieren toetsen op niveau E3 voorkomt, kan in een digitale (tussen)toets van niveau M3E3 of E3M4 voorkomen, mits de spellingcategorie ook op dat niveau bevraagd wordt. Op basis van de onderliggende schaal was het wel mogelijk om een opgave verder op te schuiven, maar wat betreft face-validity niet wenselijk.
- 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 een zelfde 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 .40 en .90) die door de betere spellers significant vaker goed worden gemaakt dan door de minder goede spellers (rir vanaf .20) in aanmerking voor opname in de definitieve digitale toetsen Spelling.

Aan de voorwaarden waarnaar deze uitgangspunten verwijzen, hebben we kunnen voldoen. De reguliere digitale toetsen (M3 en E3) hebben grote overlap (tweederde of meer) met de corresponderende reguliere papieren toetsen. De digitale tussentoets (M3E3) heeft enigszins overlap met de digitale reguliere toets E3 en met de digitale toetsen van de tweede generatie.
Als we een opgave hebben verschoven naar een ander toetsniveau ten opzichte van de papieren toetsen, is het – met uitzondering van één item – maximaal een half niveau opgeschoven. De uitzondering betreft een item dat opgenomen is in de digitale toets M3E3 en in de papieren toets E3M4.

In alle toetsen van groep 3 zijn de aantallen opgaven per categorie voor 100% gelijk aan die van de papieren variant, zie tabel 3.1.

Tabel 3.1 Spellingcategorieën in de digitale toetsen Spelling 3.0 groep 3 (met de aantallen in papieren toetsen tussen haakjes)

Cat.	Omschrijving	M3	M3E3	E3
1	mkm-woorden	14 (14)	3 (3)	
2	mmkm- en mkmm-woorden	26 (26)	6 (6)	4 (4)
3	mmkmm-woorden		7 (7)	7 (7)
4	woorden met een niet geschreven tussenklank		7 (7)	5 (5)
5	woorden met -mmm of mmm-		6 (6)	7 (7)
6	woorden met sch(r)-		5 (5)	5 (5)
7	woorden met -ng(-) of -nk(-)		6 (6)	5 (5)
8	woorden met f-, v-, s- of z-			3 (3)
9	verkleinwoord met uitgang -je(s) of -tje(s)			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 3 in relatie tot de gemiddelde vaardigheidsscore voor de afnamemomenten. 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 toets M3 zijn er veel makkelijke opgaven (die liggen onder de stippellijn van M3) en redelijk veel opgaven van gemiddelde moeilijkheid (doorkruisen de lijn van M3). Er zijn geen echt moeilijke opgaven (zouden boven de lijn van M3 moeten liggen). De toets is weliswaar makkelijk, maar heeft voldoende discriminerend vermogen: de echte probleemspellers worden gesignaleerd.
De toets E3 laat een vergelijkbaar beeld zien als de toets M3: er zijn veel makkelijke opgaven (liggen onder de stippellijn van E3) en wat moeilijkere opgaven (doorkruisen de lijn van E3 of liggen erboven).
Bij de gemakkelijke variant van de toets E3, de toets M3E3, liggen meer balkjes ónder de lijn van E3 dan bij de toets E3 (dus onder de gemiddelde vaardigheidsscore). Deze toets is over de gehele linie erg makkelijk en dus geschikt voor de zwakkere spellers.
Het beeld dat we zien bij de digitale toetsen groep 3 is vergelijkbaar met het beeld van de papieren toetsen groep 3.

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 gebruikgemaakt 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.

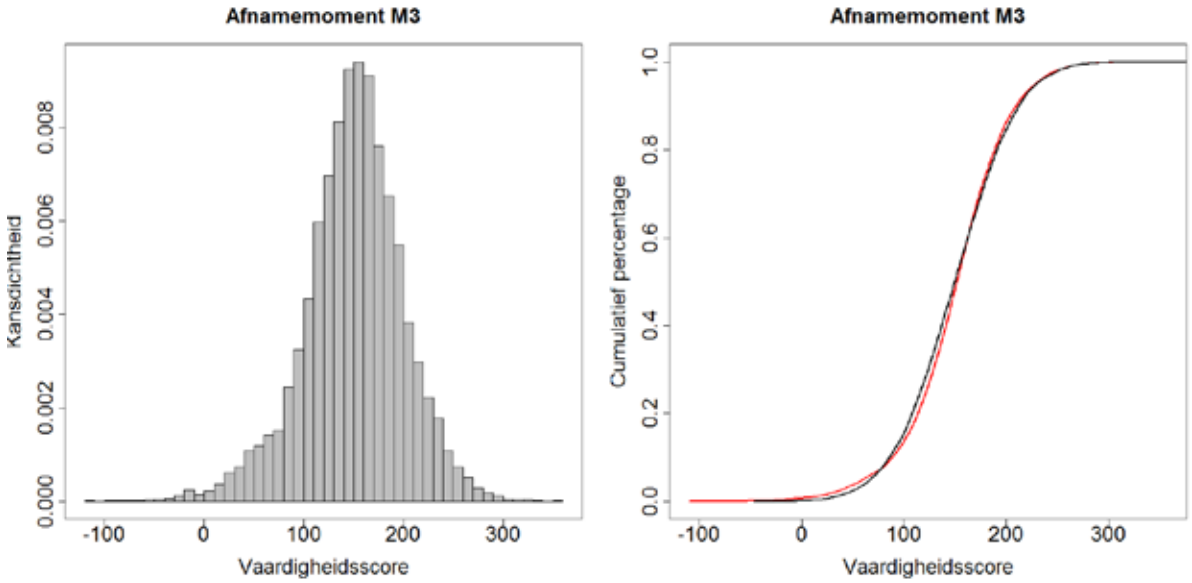
Voor de verschillende boekjes (per afnamemoment) in het kalibratieonderzoek papier-digitaal zijn de correlaties tussen de schattingen met de beide genoemde berekeningswijzen alle groter dan .99. Er zijn voor deze situatie T-testen uitgevoerd: voor alle leerlingen geldt dat er geen significant verschil 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 vaardigheids-score 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 vaardigheids-scores 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 toch de voorkeur gegeven aan de gewogen score.

In tabel 3.2 is te zien dat de parameters van de vaardigheidsverdeling van de tussentoets M3E3 gelijk is aan die van de vaardigheidsverdeling van E3. De reden is dat voor de tussentoets de normeringsgegevens van de genormeerde toets van het niveau erna (E3) zijn gebruikt. Voor tussentoets M3E3 is dus de verdeling van de vaardigheidsscore gebruikt van de E3-toets die past bij het E3-afnamemoment.
De parameters van de vaardigheidsverdelingen zijn exact hetzelfde 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 2964 leerlingen voor M3 en op 2996 leerlingen voor E3. Dit betreft de aantallen leerlingen van de normering van de papieren toetsen.
De waarden laten zien dat de vaardigheidsverdeling bij benadering normaal is. Figuur 3.1 met de verdeling van de vaardigheidsscores van M3 en figuur 3.2 met de verdeling van de vaardigheidsscores van E3 laten dit ook zien. Ook deze figuren zijn precies zo opgenomen in de wetenschappelijke verantwoording van de papieren toetsen Spelling 3.0 groep 3. Voor de duidelijkheid geven we deze figuren hier nogmaals weer.

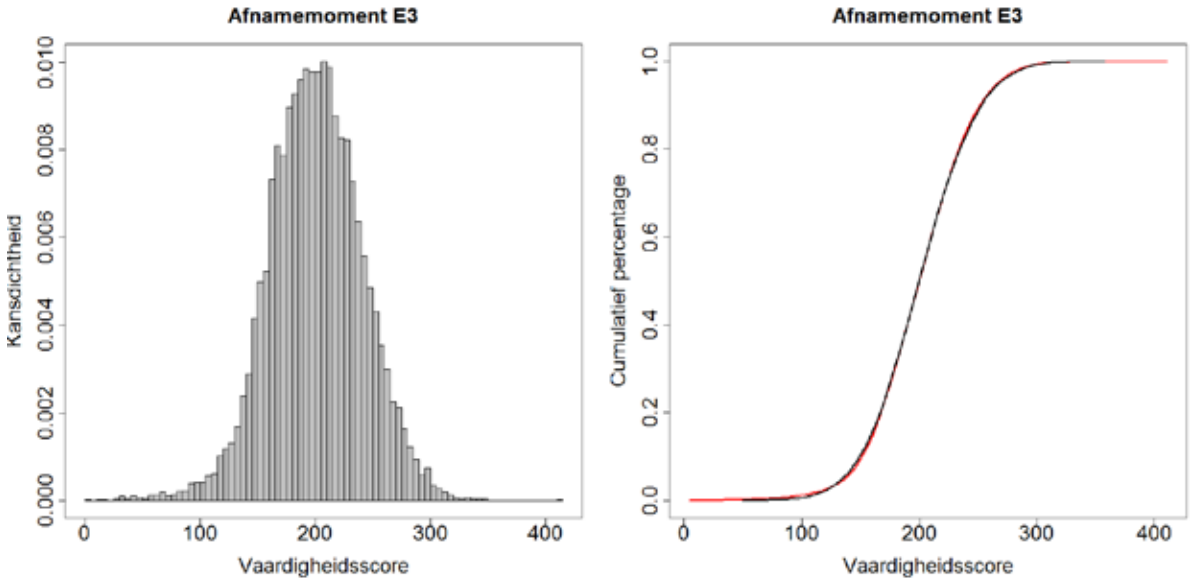
Tabel 3.2 Beschrijvende gegevens toetsen M3, M3E3 en E3 op de gewogen scoreschaal en de vaardigheidsschaal

	Gemiddelde	**Standaarddeviatie**	**Kurtosis**	**Scheefheid**
M3 Gewogen score	107,1	37,1	-0,18	-0,96
M3 Vaardigheid	150,0	50,0	1,05	-0,38
M3E3 Gewogen score	111,8	23,3	2,32	-1,56
M3E3 Vaardigheid	200,4	40,4	0,63	-0,14
E3 Gewogen score	97,5	24,6	0,67	-1,09
E3 Vaardigheid	200,4	40,4	0,63	-0,14

Figuur 3.1 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling M3 (papier)



Figuur 3.2 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling E3



4 Kalibratie en normering

4.1 Rationale van de kalibratie-onderzoeken

Aan het begin van het toetsontwikkelingsproces van de LVS-toetsen Spelling 3.0 groep 3, zijn in 2011 opgaven geconstrueerd. Deze opgaven zijn allereerst in *papieren* versie onderzocht in kalibratie-onderzoeken in 2012 en – na opname van de opgaven in de uit te geven 'papieren' toetsen – in normeringsonderzoeken in 2013. Zie voor het kalibratie- en normeringsonderzoek van de papieren items voor groep 3 paragraaf 4.2 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling voor groep 3 (Tomesen, Wouda, Mols & Horsels, 2015).

In 2014 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? Indien dit het geval is, passen de papieren en digitale items op dezelfde schaal.
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 (bijvoorbeeld M3) 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 zal worden besproken hoe de digitale en papieren opgaven op de vaardigheidsschaal spelling passen. De papieren en de digitale opgaven meten met andere woorden dezelfde vaardigheid. 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 3 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 passen. In januari 2014 vond het papier-digitaal-kalibratieonderzoek M3 plaats. In juni 2014 vond het papier-digitaal-kalibratieonderzoek E3 plaats. In deze onderzoeken werd slechts een deel van de items meegenomen die geselecteerd waren voor de papieren uitgaven 3.0. Het was onderzoekstechnisch niet mogelijk om alle papieren items mee te nemen. Om alle items te kunnen onderzoeken, zouden grote

aantallen scholen en leerlingen nodig zijn geweest. En we wisten op basis van ervaringen met het werven van scholen voor andere proeftoetsen die digitaal werden afgenomen, dat de bereidheid onder scholen om deel te nemen aan dergelijke digitale onderzoeken zeer laag was. Om die reden hebben we noodgedwongen gekozen voor een opzet en design waarbij we zo min mogelijk leerlingen nodig zouden hebben, maar waarbij nog wel toetsen van goede kwaliteit samengesteld konden worden. De opgaven van de reguliere papieren toetsen 3.0 M3 en E3 zijn omgezet naar digitale versies. De digitale tussentoets M3E3 is slechts gedeeltelijk omgezet naar een digitale versie. Voor de uitgave zouden we vervolgens naast nieuwe opgaven ook opgaven uit de digitale Starttaak LVS Spelling tweede generatie (die eveneens in de kalibratieonderzoeken werd meegenomen) selecteren.

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 tabel 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.
De kalibratieonderzoeken voor de twee verschillende afnamemomenten kennen elk een iets ander design. Daarom bespreken we hieronder opzet en design voor elk kalibratieonderzoek afzonderlijk.

Medio 3
In het papier-digitaal onderzoek voor het 'medio' (M) afnamemoment van januari 2014 zijn 85 items voorgelegd aan 414 leerlingen van groep 3. Elke leerling maakte een taak op papier en een taak digitaal. De items waren verdeeld over twee verschillende opgavenboekjes in een onvolledig, maar via het normeringsonderzoek van de papieren toets 'verbonden' design. De taken 1 en 2 bestonden uit gedigitaliseerde items van de papieren toets M3 3.0. Elke leerling maakte één digitale taak met 30 nieuwe items uit LVS Spelling 3.0. Daarnaast maakte elke leerling de Starttaak M3 van LVS tweede generatie op papier. Het was voor de leerling niet mogelijk om deze taak digitaal te maken, omdat er geen digitale toets M3 van de tweede generatie is.

De data uit dit papier-digitaal onderzoek werden vervolgens gekoppeld aan de data die werden verzameld in het normeringsonderzoek voor de papieren uitgave M3 uit 2013. 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 M3

	Taak			**Aantal leerlingen**
	Papier	**Digitaal**		
Boekje	Starttaak M3 LVS 2e generatie	Taak 1 3.0	Taak 2 3.0	
1				206
2				208
Aantallen per taak	414	206	208	

Eind 3

In het papier-digitaal onderzoek voor het 'einde' (E) afnamemoment van juni 2014 zijn 140 items voorgelegd aan 623 leerlingen van groep 3. Elke leerling maakte één digitale taak met 30 nieuwe items uit LVS Spelling 3.0. De taken 1 en 2 bestonden uit gedigitaliseerde items van de papieren toets E3 3.0. Taak 3 bestond uit gedigitaliseerde items van de papieren tussentoets M3E3 3.0[1], die net als de toets E3 uit zinsdictee bestaat. Daarnaast maakte elke leerling de Starttaak E3 van LVS tweede generatie. Ongeveer de helft van de leerlingen maakte de papieren Starttaak van de tweede generatie (de leerlingen die toegewezen waren aan boekje 1, 2 en 3) en ongeveer de helft maakte de digitale Starttaak van de tweede generatie (de leerlingen die toegewezen waren aan boekje 4, 5 en 6). 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 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 zes 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 E3 uit 2013. 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 E3

	Taak					**Aantal leerlingen**
	Papier	**Digitaal**				
Boekje	Starttaak E3 LVS 2e generatie	Starttaak E3 LVS 2e generatie	Taak 1 3.0	Taak 2 3.0	Taak 3 3.0	
1	■		■			127
2	■			■		134
3	■				■	71
4		■	■			68
5		■		■		101
6		■			■	122
Aantallen per taak	332	291	195	235	193	

Lage aantallen, voorlopige normering en aanvullende dataverzameling via dataretour

Uit de afgenomen taken werden de digitale toetsen M3, M3E3 en E3 samengesteld. Hierbij bleek dat niet alle items door een toereikend aantal leerlingen waren gemaakt om ze met voldoende precisie te kunnen kalibreren en om vast te stellen of er sprake was van differentieel itemfunctioneren (DIF) ten opzichte van de papieren items. Vanwege de ontoereikendheid van de data was er onvoldoende vertrouwen in het zonder meer kunnen toepassen van de 'papieren normering' op de digitale toetsversies.

[1] Voor de digitale tussentoets M3E3 konden we geen gebruik maken van opgaven uit de toets M3, omdat de toets M3 uit opgaven woorddictee bestaat en de tussentoets M3E3 uit opgaven zinsdictee. Daarom hebben we in het kalibratie-onderzoek E3 ook een gedigitaliseerde tussentaak M3E3 moeten inzetten. (In de kalibratieonderzoeken in groep 4 en verder zijn geen gedigitaliseerde tussentaken meer ingezet.)

Daarom is ervoor gekozen om de toetsen uit te geven met voorlopige normering en de toetsen opnieuw te kalibreren nadat er nieuwe data verzameld waren door middel van dataretour. Dataretour is de optie waarbij scholen Cito toestemming geven om de toetsgegevens van leerlingen te gebruiken voor toetsconstructie-, normerings- en onderzoeksdoeleinden. Op deze manier was het mogelijk via digitaal ingestuurde afnames van de nieuwe digitale toetsen voor voldoende aantallen leerlingen te zorgen, zodat een betrouwbare (nieuwe) kalibratie kon worden uitgevoerd.

Uiteindelijke aantallen in de kalibratieanalyses
De in de volgende paragrafen beschreven kalibratie is gebaseerd op de minimale aantallen afnamen per item (item dat het minst vaak gemaakt is) zoals deze in tabel 4.3 worden vermeld. De aanvullende afnamen door middel van dataretour zijn gedurende 2016 gerealiseerd en door scholen ingestuurd. De tabel laat zien dat de aantallen via dataretour gerealiseerde afnames sterk verschillen. Met name de tussentoets wordt relatief weinig afgenomen, wat gezien de functie van deze toets begrijpelijk is. De totale aantallen afnames die in de analyses konden worden meegenomen volstonden echter voor een nauwkeurige kalibratie (zie de laatste kolom van tabel 4.3).

Tabel 4.3 Minimale aantallen afnamen per item van het kalibratie-onderzoek voor de toetsen Spelling 3.0 M3, M3E3 en E3

Toets	**Aantal afnamen steekproef (minimaal)**	**Aantallen afnamen aanvulling dataretour**	**Aantallen kalibratie-onderzoek (minimaal)**
M3	206	835	1041
M3E3	194	120	314
E3	194	1044	1238

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

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

Deze 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 overschrijdingskansen uniform verdeeld moeten zijn op het (0,1) interval.
Kortom, als we afzien van de formeel-statistische achtergrond van de gehanteerde toetsen, kan de kalibratieprocedure als volgt worden samengevat:

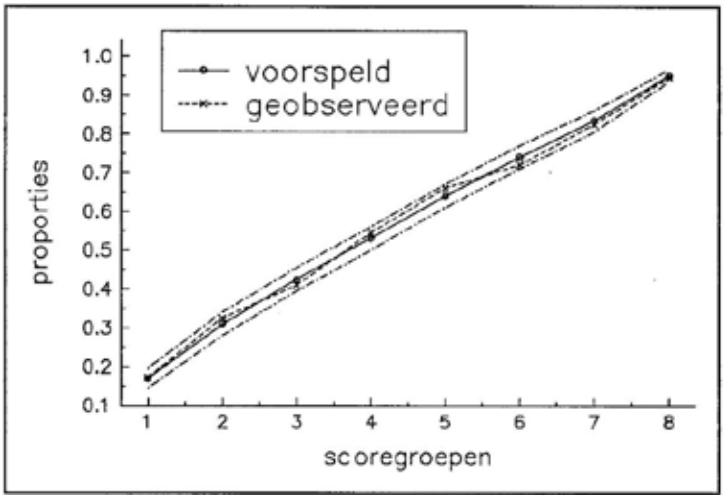
1 Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
2 Vervolgens schatten we de itemparameters met behulp van de CML-methode.
3 Met behulp van de M-toetsen controleren we of de discriminatie-indices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

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

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 4.1 (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 3 zelfs bij de slechtst passende opgave sprake is van een zeer aanvaardbaar beeld. Er wordt in dit geval voor een deel (van de onderscheiden scoregroepen) niet beantwoord aan de eis dat de geobserveerde proportie binnen het 95%-betrouwbaarheidsinterval van de geschatte proporties ligt. Dit beeld doet zich slechts bij enkele opgaven voor die dan ook een uitzondering vormen. De overige opgaven voldoen voor alle scoregroepen wel aan die eis. De afbeeldingen voor de representatieve en best passende opgaven illustreren dit. Dit leidt tot de conclusie dat bij vrijwel alle opgaven in de toetsen Spelling een grafische voorstelling van de S-toetsing hoort die in grote lijnen met figuur 4.1 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 3. 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).

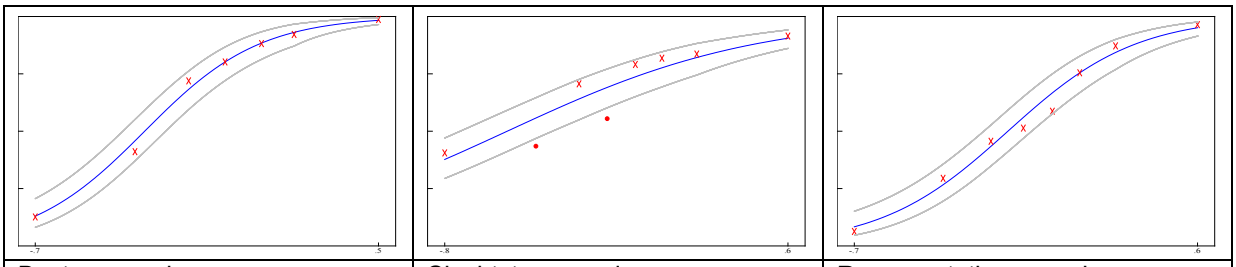
4.2.4 Totale kalibratie per groep

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. Dat wil zeggen dat voor alle items van de drie toetsen van groep 3 een kalibratie werd uitgevoerd. Hierbij werden de parameters van de op papier afgenomen items gefixeerd op de waarden zoals geschat in de normerings-onderzoeken 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. Hieronder zal daarom ook de passing van het meetmodel van de totale kalibratie per groep besproken worden.

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.

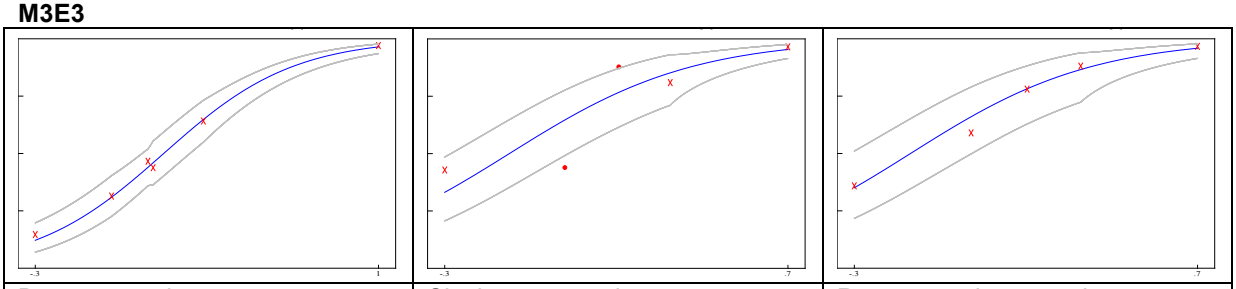
Figuur 4.2 Voorbeelden van S-toetsen voor de digitale toetsen Spelling 3.0 M3, M3E3 en E3 met per toets de best passende, de slechtst passende en een qua passing representatieve opgave

M3



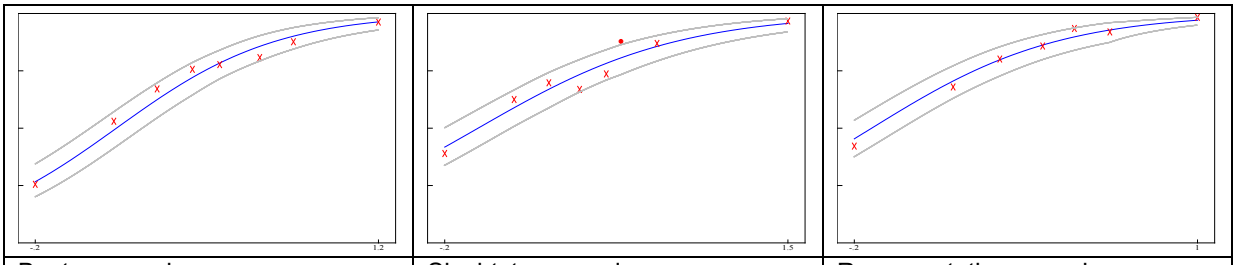
Best passend	Slechtst passend	Representatieve passing

M3E3



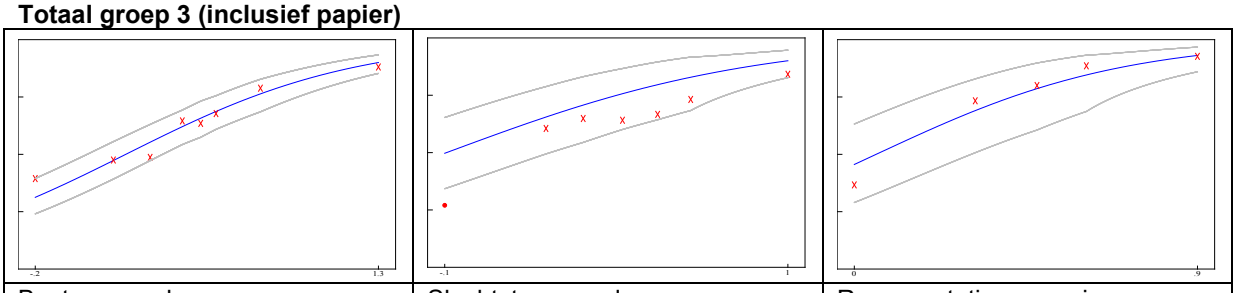
Best passend	Slechtst passend	Representatieve passing

E3



Best passend	Slechtst passend	Representatieve passing

Totaal groep 3 (inclusief papier)



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 4.4 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 3. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan .01, respectievelijk .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 4.4 Verdeling van overschrijdingskansen bij S-toetsen voor digitale toetsen Spelling 3.0 groep 3 en bij de totale kalibratie.

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
M3	2	5	2	4	6	4	1	4	4	3	2	3
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
M3E3	1	2	2	2	2	3	1	3	3	7	5	9
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
E3	0	1	3	3	6	3	4	2	3	3	5	7
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
G. 3 Tot	6	11	22	39	38	29	25	29	21	38	39	31

In tabel 4.5 zijn de R1c-waarden weergegeven voor dezelfde afnames waarvoor in tabel 4.4 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 4.5 blijkt dat de modelpassing voor alle toetsen goed is. De significantie van de statistische toetsingen is bij de grote aantallen in de analyse (tussen duizend en tweeduizend voor groep 3) nauwelijks informatief.

Opgemerkt moet worden dat de R1c waarden niet alleen van toepassing zijn op de toetsversies zoals ze uiteindelijk zijn samengesteld; de niet-geselecteerde items zijn (om de verbondenheid van het design recht te doen) ook meegenomen in de analyses. Het ligt voor de hand dat deze aanpak de modelfit negatief heeft beïnvloed, want in de regel vallen minder goed passende items uit de itembank bij de selectie af.
Dit impliceert dat de R1c-toetsingen (nog) gunstiger zouden zijn uitgevallen wanneer uitsluitend geselecteerde items zouden zijn meegenomen (NB. Zwakkere, niet geselecteerde items hebben een relatief hoge bijdrage aan de toetsingsgrootheid R1c).

Tabel 4.5 R1c-waarden voor de digitale toetsen Spelling 3.0 M3, M3E3 en E3

Toetsversie	**R1c**	**df**	**p**
M3	403,794	308	0,0001
M3E3	367,774	316	0,0227
E3	454,345	336	<0,00005
Groep 4 totaal	2643,276	2077	<0,00005

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 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.
De waarden voor deze constante zijn weergegeven in tabel 4.6. De gemiddelde waarden van de constante zijn goed te noemen. Eén item (uit M3E3) had een waarde boven de 0,2, maar deze waarde lag nog onder de 0,3.

Voor de totale kalibratie van groep 3 lagen 13 waarden boven de 0,2 en 11 van deze waarden lagen onder de 0,3. Het ene item met een waarde van 'c' die door de COTAN als 'matig' wordt omschreven is niet in een van de toetsen opgenomen.

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

Tabel 4.6 Nauwkeurigheid van de itemparameterschattingen (constante 'c')

Toetsversie	Constante 'c'	
	Range	Gemiddelde
M3	0,031 – 0,123	0,055
M3E3	0,045 – 0,212	0,105
E3	0,043 – 0,103	0,064
Groep 3 totaal	0,029 – 0,489	0,107

Ook op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de digitale toetsen Spelling 3.0 M3, M3E3 en E3 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 nauwelijks tot geen sprake is van DIF.
Een enkel digitaal item kon niet op dezelfde schaal worden gebracht en is dan ook niet in de digitale toetsversie opgenomen. 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 de afnametijdstippen M3 en E3. 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 (M3 en E3) op deze schaal gebruiken. Voor het beschrijven van de normpopulatie kunnen daarom de eerder gerapporteerde resultaten gebruikt worden.

De Expertgroep Toetsen PO had als oordeel dat voor de normering van Spelling 3.0 groep 3 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 LVS-toetsen Spelling 3.0 voor groep 3 (Tomesen, Wouda, Mols & Horsels, 2015).

Sinds schooljaar 2013/2014 wordt door Cito een nieuwe werkwijze voor het normeren van leerling-

volgsysteemtoetsen toegepast. Deze werkwijze wordt gebruikt bij het monitoren van de normering van inmiddels uitgegeven toetsen, maar wordt ook gebruikt bij de normering van de nieuw uit te geven toetsen, zo ook bij de derde generatie toetsen voor Spelling. De werkwijze die we hieronder beschrijven, komt uit Keuning et al. (2014). Allereerst besteden we aandacht aan de opzet van het normerings-onderzoek, de gehanteerde procedures en de aantallen leerlingen per afnamemoment (paragraaf 4.3.1). Vervolgens komt in paragraaf 4.3.2 de representativiteit van de normsteekproeven aan de orde. De paragraaf wordt afgerond met een presentatie van de resultaten van de normering (i.e. de kenmerken van de 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 van de wetenschappelijke verantwoording van de papieren toetsen Spelling 3.0 voor groep 3 (Tomesen, Wouda, Mols & Horsels, 2015)) werden data verzameld. Om deelnemers te werven voor het normeringsonderzoek zijn scholen aangeschreven. Voor het embedded field normeringsonderzoek is een representatieve steekproef getrokken uit de verzameling van alle basisscholen in Nederland. Dit is gedaan vanuit het bij Cito gebruikelijke steekproefkader dat bepaald wordt door regio, urbanisatiegraad en schooltype (zie verderop voor een omschrijving van deze achtergrondvariabelen).

Voor het normeringsonderzoek van de *toets M3* waren 1200 leerlingen nodig. Er zijn 1221 scholen aangeschreven. Hiervan waren 68 scholen met in totaal 1660 leerlingen bereid om deel te nemen. Uiteindelijk hebben 65 scholen met 1518 leerlingen daadwerkelijk meegedaan en konden we de gegevens van 1482 leerlingen gebruiken. Leerlingen met veel missings werden namelijk niet meegenomen bij de kalibratie/normering.
De normeringsgroep voor de *toets E3* bestond deels uit herhalingsscholen die ook op het afnamemoment M3 en/of M4 hadden meegedaan aan het normeringsonderzoek. (De meeste scholen deden aan zowel M3 als M4 mee.) Voor het normeringsonderzoek E3 zijn uiteindelijk in totaal 77 herhalingsscholen van M3/M4 aangeschreven en 854 extra scholen. Omdat na de eerste aanschrijvingsronde maar 49% van de herhalingsscholen uit het normeringsonderzoek Spelling M3/M4 ook bereid bleek deel te nemen aan het normeringsonderzoek Spelling E3 en minder dan 1% van de scholen uit de aanvullende steekproef bereid bleek deel te nemen aan het normeringsonderzoek zijn in een tweede wervingsronde 1400 extra scholen aangeschreven. In totaal meldden zich 69 scholen aan voor het normeringsonderzoek E3 met in totaal 1588 leerlingen. Uiteindelijk hebben 67 scholen met 1525 leerlingen daadwerkelijk meegedaan en konden we de gegevens van 1493 leerlingen gebruiken.

Voor het bepalen van de normering werden de gegevens uit het normeringsonderzoek aangevuld met gegevens uit Cito dataretour. In tabel 4.7 zijn de uiteindelijke aantallen scholen en leerlingen in het (‘papieren’) normeringsonderzoek samengevat.

Tabel 4.7 Aantal leerlingen per afnamemoment die meegenomen zijn in de normering

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

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 populatie-verdelingen 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 zeer goede afspiegeling vormen van de populatie.

4.3.3 Normeringsresultaten

Na de hierboven beschreven procedure doorlopen te hebben en de normeringssteekproef te hebben samengesteld, kon de normering worden bepaald. Naast het gemiddelde werden de percentielen bepaald. Dat gebeurde op basis van de verdeling van scores die werden gevonden in de normeringssteekproef zoals die is samengesteld op basis van het embedded field normeringsonderzoek en Cito-dataretour. Om de scores van leerlingen te kunnen vergelijken over de tijd worden vaardigheidsscores gebruikt. Uit de ruwe scores van de leerlingen uit het embedded field normeringsonderzoek en Cito-dataretour werden “plausible values” gegenereerd op de nieuw ontwikkelde vaardigheidsschaal. Deze “plausible values” representeren de volledige range aan vaardigheidsscores die een leerling zou kunnen hebben, gegeven de scores.
De “plausible values” geven niet alleen informatie over de geschatte vaardigheid maar ook over de onzekerheid die bij die schatting hoort (Keuning et al., 2014). De normering werd vervolgens gebaseerd op de “plausible values” van de leerlingen in de normeringssteekproef. Tabel 4.8 geeft de normgegevens voor de toetsen Spelling 3.0 groep 3.

Tabel 4.8 Normtabel op leerlingniveau voor Spelling 3.0 groep 3

Tijd	***M***	***SD***	***K***	***S***	***P*10**	***P*20**	***P*25**	***P*40**	***P*50**	***P*60**	***P*75**	***P*80**
M3	150,0	50,0	1,045	-0,380	89,0	113,8	122,0	141,4	152,1	162,6	181,4	188,9
E3	200,4	40,4	0,631	-0,142	151,6	167,6	173,8	190,2	200,2	210,4	227,0	233,8

De betreffende normeringstabel is niet alleen van toepassing op de papieren versie van de toetsen Spelling 3.0 voor groep 3, 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 OPTAL (Verstralen, 1997).

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'.

Tabel 5.1 bevat informatie over de meeteigenschappen van de vaardigheidsschaal Spelling, voor de digitale toetsen voor groep 3. In de eerste kolom staat de aanduiding van de toets. 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 scores van de leerlingen op de verschillende toetsen.
De vijfde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van iedere toets.
De zesde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de verschillende toetsen is.

De betrouwbaarheidscoëfficiënten zijn zonder uitzondering hoog. Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen Spelling 3.0 groep 3) geeft de

COTAN aan dat een betrouwbaarheidscoëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheids-coëfficiënt 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 alle toetsen goed te noemen.

Tabel 5.1 Beschrijvende gegevens bij de digitale toetsen Spelling 3.0

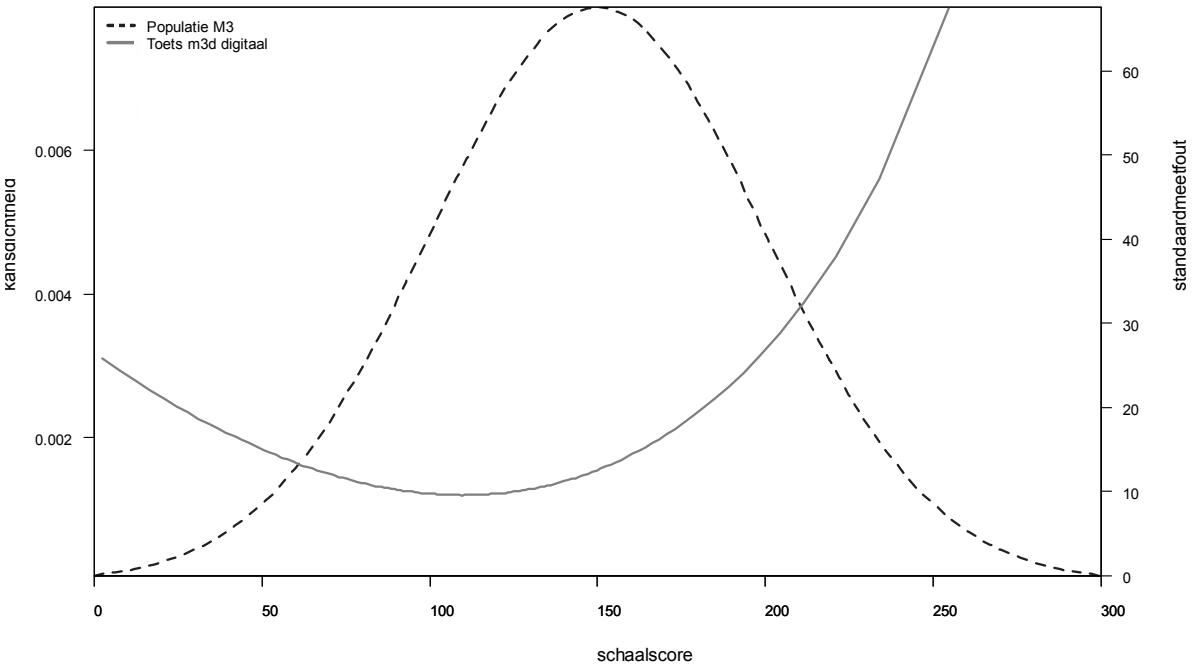
Toets	Aantal items	Maximum score	Gemiddelde	Standaard-meetfout	MAcc	Test-hertest (simulatie)
M3	40	146	107,1	8,39	0,95	0,95
M3E3	40	135	111,8	6,97	0,91	0,91
E3	40	127	97,5	7,31	0,91	0,91

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 echter OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1000000 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 1000000 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 tabel 5.1. De uitkomsten komen 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.

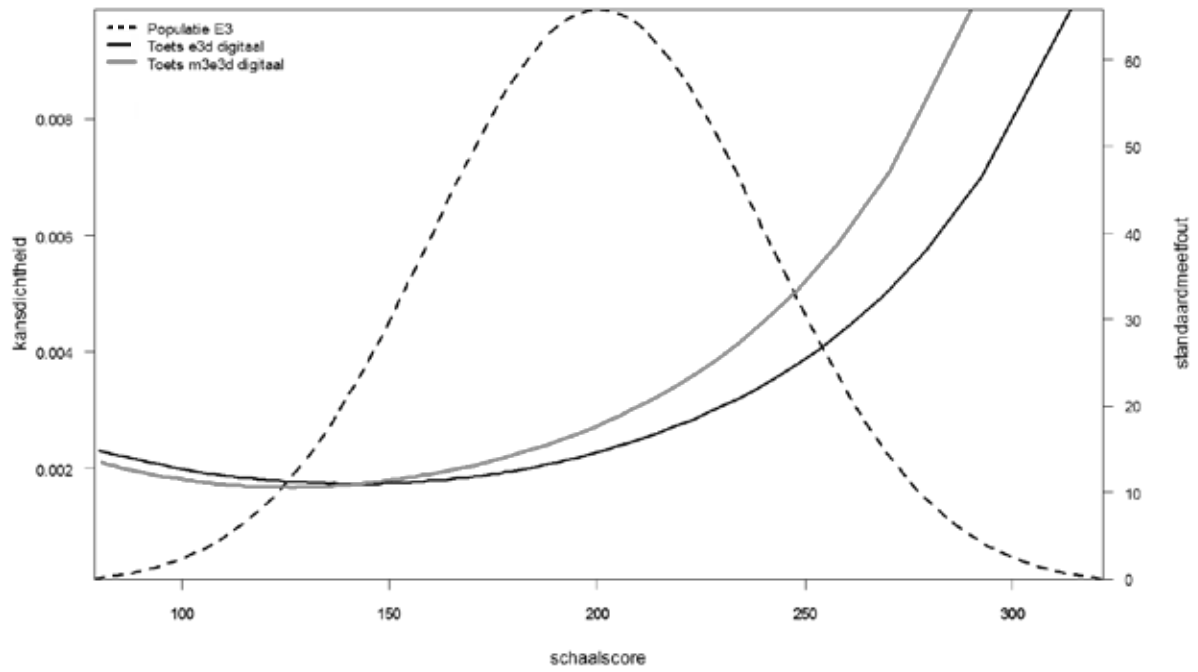
5.2 Nauwkeurigheid

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 5.1 en 5.2 geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid van de verschillende toetsen. In deze figuren staat voor iedere toets de grootte van de meetfout op de vaardigheidsschaal afgebeeld (met verdelingskenmerken zoals aangegeven in tabel 4.8). Ook zijn de kansdichtheidfuncties voor de normgroepen op de verschillende afnamemomenten 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 5.1 Grootte van de meetfouten voor de digitale toets M3 en de kansdichtheidsfunctie voor de M3-populatie



Figuur 5.2 Grootte van de meetfouten voor de digitale toetsen M3E3 en E3 en de kansdichtheidsfunctie voor de E3-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 tot en met 5.4 laten voor de digitale toetsen M3, M3E3 en E3, respectievelijk afnamemomenten medio groep 3 en einde groep 3, zien hoe vaak de werkelijke vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat

tabel 5.2 zien dat 88,1 procent van de leerlingen die halverwege groep 3 op basis van de digitale M3-toets in scoregroep V geclassificeerd wordt ook met hun werkelijke vaardigheidsscore in deze scoregroep geclassificeerd wordt. De kans dat een V-leerling terecht als V-leerling wordt bestempeld is, met andere woorden, ongeveer 88 procent. Verder laat de linkerkant van tabel 5.2 zien dat 9,0 procent van de leerlingen in scoregroep V een vaardigheidsscore heeft die in werkelijkheid in scoregroep IV valt.
De overige getallen in tabellen 5.2 tot en met 5.4 zijn op dezelfde wijze te interpreteren.

Tabel 5.2 Betrouwbaarheidstabel Toets M3 digitaal voor afnamemoment medio 3

	Scoregroep waarin de ware score valt						**Scoregroep waarin de ware score valt**				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	88,1	9,0	0,0	0,0	0,0	E	84,5	8,6	0,0	0,0	0,0
IV	11,9	71,8	13,2	0,3	0,0	D	15,5	73,7	8,3	0,0	0,0
III	0,0	18,9	60,4	17,0	1,3	C	0,0	17,7	73,2	12,6	0,2
II	0,0	0,3	25,2	56,6	20,9	B	0,0	0,0	18,3	65,0	16,9
I	0,0	0,0	1,2	26,1	77,8	A	0,0	0,0	0,2	22,4	82,9

Tabel 5.3 Betrouwbaarheidstabel Toets M3E3 digitaal voor afnamemoment einde 3

	Scoregroep waarin de ware score valt						**Scoregroep waarin de ware score valt**				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	78,5	11,2	0,6	0,0	0,0	E	77,3	9,3	0,1	0,0	0,0
IV	20,2	52,9	18,2	3,4	0,3	D	22,0	58,2	11,8	0,6	0,0
III	1,3	29,9	44,3	22,1	4,7	C	0,7	31,2	57,0	20,1	2,9
II	0,0	5,8	31,0	43,9	24,0	B	0,0	1,3	28,2	52,5	24,9
I	0,0	0,3	5,9	30,5	71,0	A	0,0	0,0	2,9	26,8	72,2

Tabel 5.4 Betrouwbaarheidstabel Toets E3 digitaal voor afnamemoment einde 3

	Scoregroep waarin de ware score valt						**Scoregroep waarin de ware score valt**				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	80,7	12,8	0,5	0,0	0,0	E	75,6	9,5	0,1	0,0	0,0
IV	18,5	57,6	19,4	2,3	0,1	D	23,6	60,0	11,7	0,3	0,0
III	0,8	26,4	48,8	22,1	3,0	C	0,8	29,8	60,4	18,0	1,3
II	0,0	3,1	28,1	48,6	22,5	B	0,0	0,8	26,2	56,5	20,9
I	0,0	0,0	3,2	27,0	74,4	A	0,0	0,0	1,5	25,2	77,8

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

De samenvattende indices voor afnamemomenten medio groep 3 en einde groep 3 zijn te vinden in tabel 5.5. Waar de betrouwbaarheidstabellen laten zien dat de meeste leerlingen op basis van hun geschatte vaardigheidsscore geplaatst worden in de niveaugroep waar ze werkelijk thuishoren, maakt tabel 5.5 aannemelijk dat de uitkomsten duidelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969) of zelfs boven dit ambitieniveau uitstijgen. Gemiddeld gezien scoort, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 95,5 tot 99,9 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 58,1 tot 75,9 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij beide afnamemomenten gemiddeld gezien in zo'n 68 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. De resultaten stemmen hiermee tot grote tevredenheid: het percentage misclassificaties is erg beperkt. De laagste waarden zien we bij toets M3E3, en dan met name bij de hoogste scoregroep. Dat is conform verwachting, aangezien deze toets – die wat makkelijker is dan de toets E3 – expliciet bedoeld is voor de minst vaardige leerlingen einde groep 3. De (boven)gemiddeld vaardige leerlingen zullen deze toets in de praktijk ook niet maken.
Op basis van bovenstaande gegevens concluderen we dat op basis van de digitale toetsen Spelling 3.0 groep 3 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 1 niveau verschil.

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

	ToetsM3 digitaal, afnamemoment M3		**Toets M3E3 digitaal, afnamemoment E3**		**Toets E3 digitaal, afnamemoment E3**	
	score-groep I t/m V	**score-groep A t/m E**	**score-groep I t/m V**	**score-groep A t/m E**	**score-groep I t/m V**	**score-groep A t/m E**
Marginal classification accuracy	70,9	75,9	58,1	63,4	62,0	66,1
Accuracy plus/minus 1 niveau	99,4	99,9	95,5	98,3	97,4	99,0

6 Validiteit

Voor de verantwoording van de validiteit verwijzen we naar hoofdstuk 6 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling groep 3 (Tomesen, Wouda, Mols & Horsels, 2015). Alles wat beschreven staat in dit hoofdstuk gaat ook op voor de digitale toetsen Spelling groep 3. Daarbij geldt dat ook de modelfit voor de digitale opgaven goed 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 wordt kort weergegeven wat in de voorafgaande hoofdstukken is besproken.
In hoofdstuk 1 is aangegeven dat hier om een *aanvulling* gaat bij de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3 (Tomesen, Wouda, Mols & Horsels, 2015). Deze aanvulling heeft uitsluitend betrekking op de *digitale* toetsen Spelling 3.0 voor groep 3.
Net als de papieren LVS-toetsen Spelling 3.0 groep 3 vormen de digitale LVS-toetsen Spelling 3.0 voor groep 3 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 verwezen we naar de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3. Meetpretentie, gebruiksdoel en functie zijn identiek voor de papieren en digitale toetsen. Een klein verschil met betrekking tot de doelgroep lichtten we toe.
Specifieke uitgangspunten bij het samenstellen van de digitale toetsen werden in hoofdstuk 3 beschreven. Dit hoofdstuk bevat ook een beschrijving van enkele psychometrische kenmerken.

In hoofdstuk 4 is verantwoord over de kalibratie- en normeringsonderzoeken. De kalibratie-onderzoeken gebeurden in de vorm van papier-digitaal vergelijkingsonderzoeken. Hieruit bleek dat de digitale items op dezelfde schaal pasten en daardoor dezelfde vaardigheid meten als de papieren items. De modelfit voor de digitale items is uitstekend en daarmee is voldaan aan de eisen van unidimensionaliteit.
De normering lag al vast voor de papieren items. Deze hebben we ook aangehouden voor de digitale items.

In hoofdstuk 5 werd 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 3 zeer hoog. Ze variëren van 0,91 tot 0,95. Verder zijn in dit hoofdstuk betrouwbaarheidstabellen opgenomen die de betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden, laten zien. Daarnaast gaven we inzicht in de lokale betrouwbaarheid: de meetfout blijkt het kleinst te zijn in de lagere en gemiddelde vaardigheidsregionen.

Voor de verantwoording van de validiteit (hoofdstuk 6) is weer verwezen naar de Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 groep 3.

Aanvullende literatuur

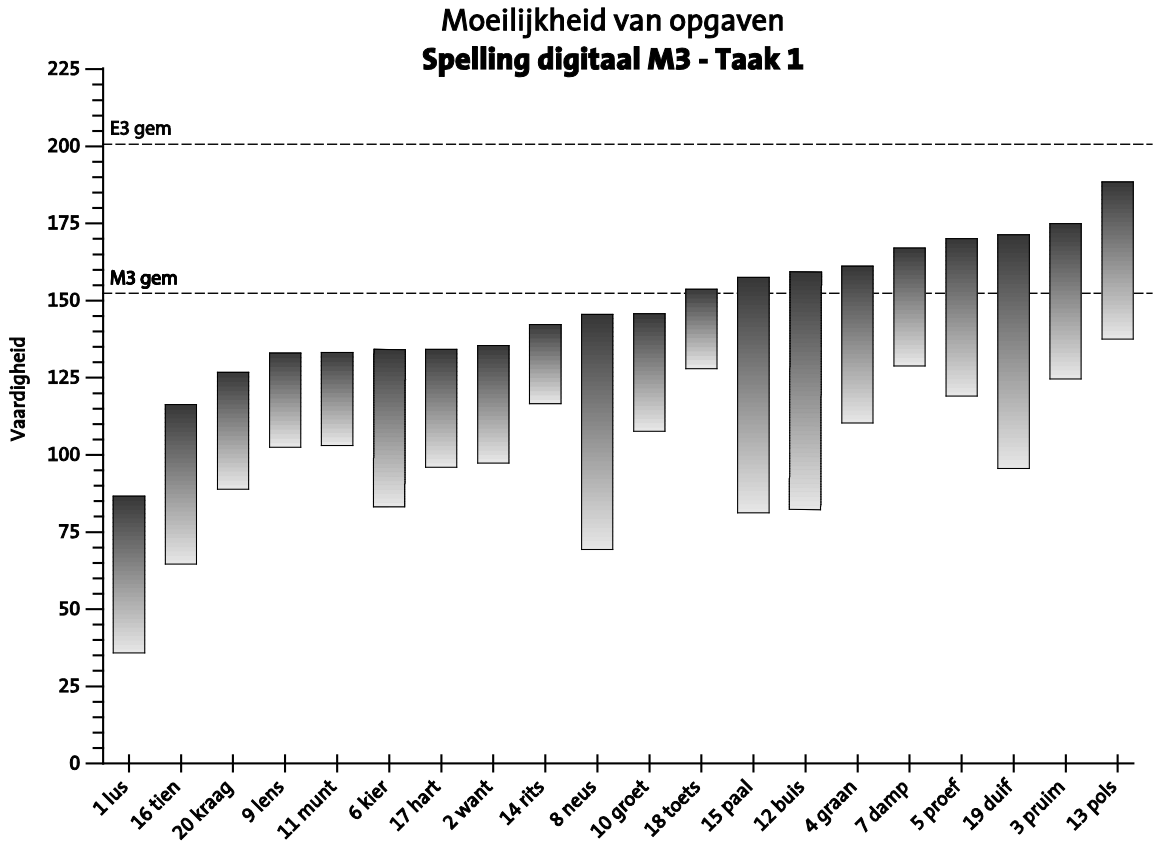
Cito (2014). *Cito Volgsysteem primair en speciaal onderwijs. Spelling 3.0 Groep 3*. Arnhem: Cito.

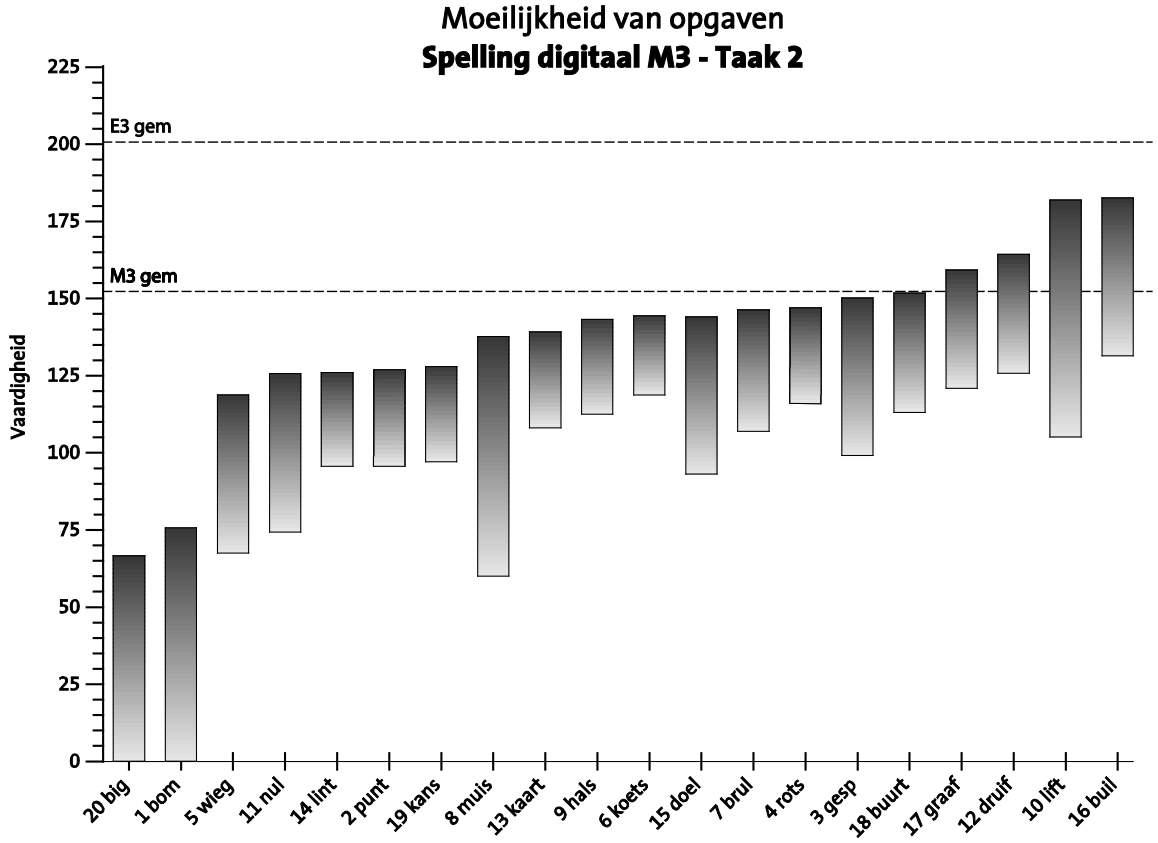
Cito (2016). *Spelling 3.0 Handleiding digitale toetsen*. Arnhem: Cito.

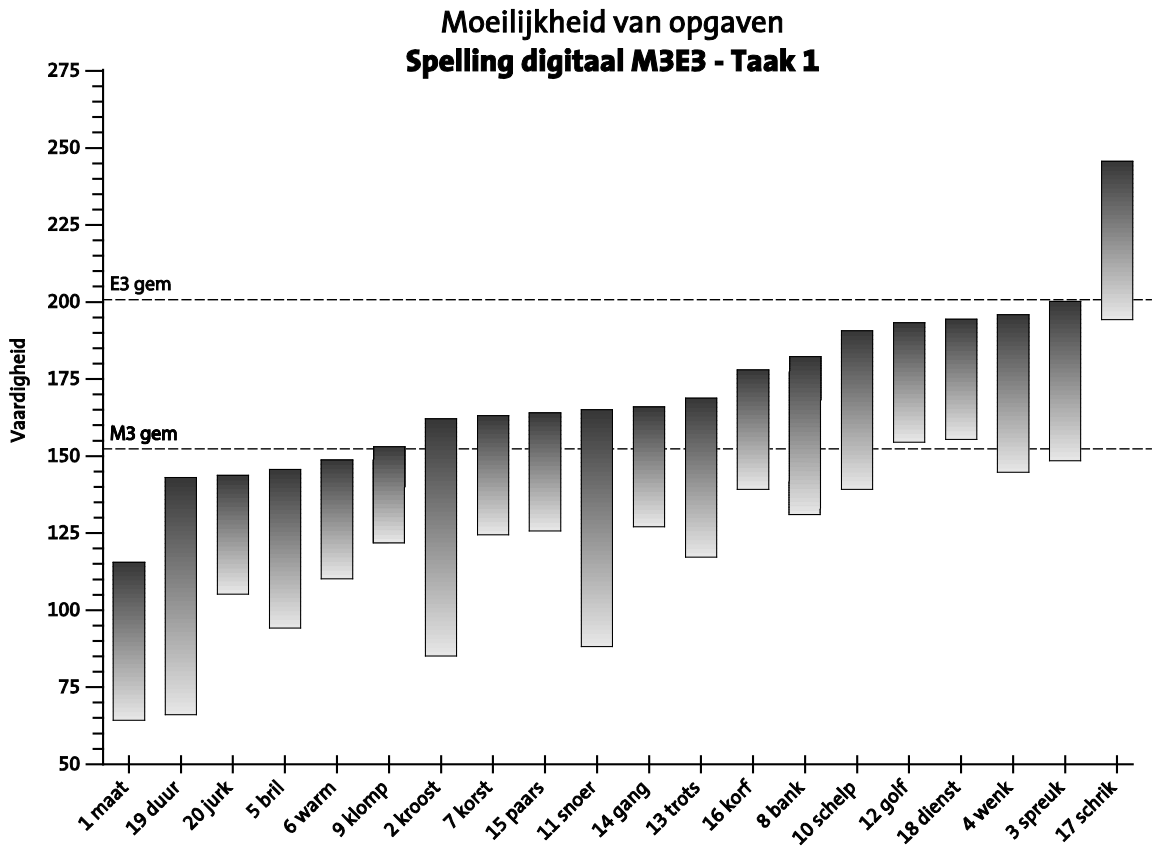
Tomesen, M., Wouda, J., Mols, A. & Horsels, L. (2015). *Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 3*. Arnhem: Cito.

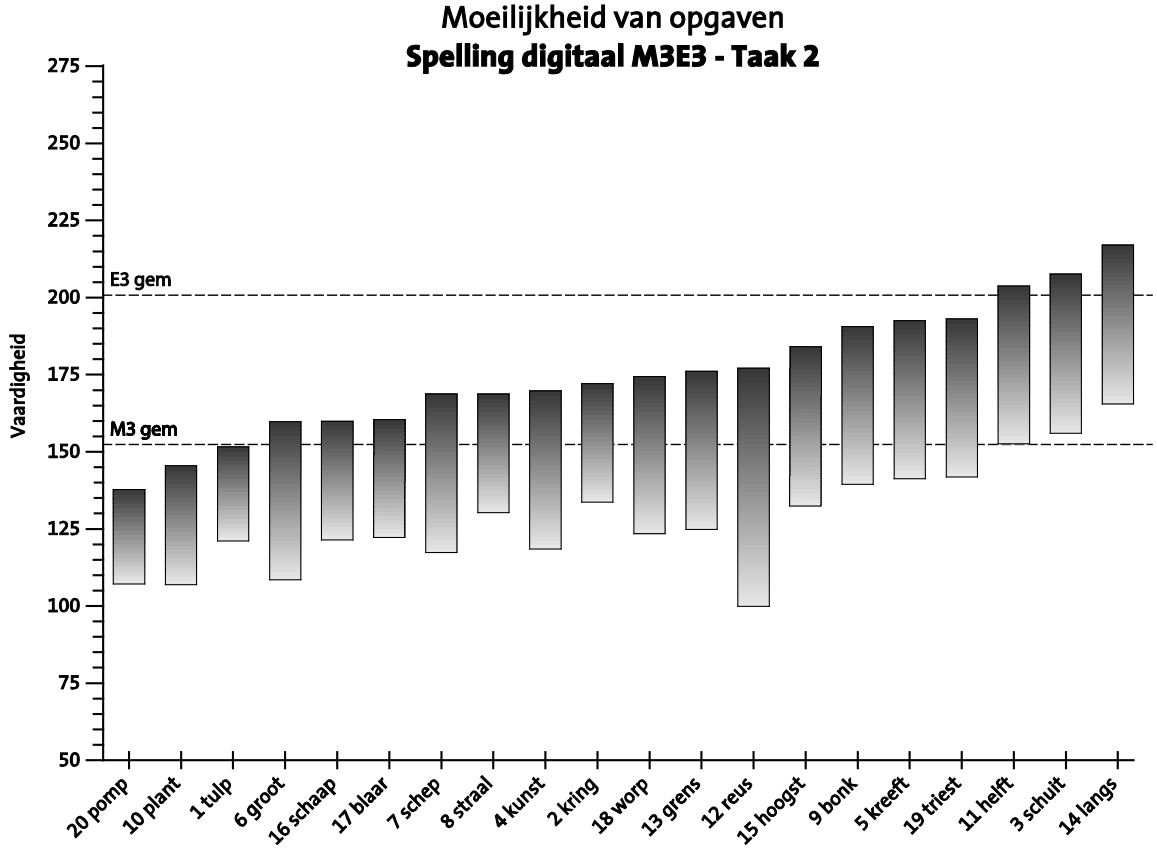
Bijlagen

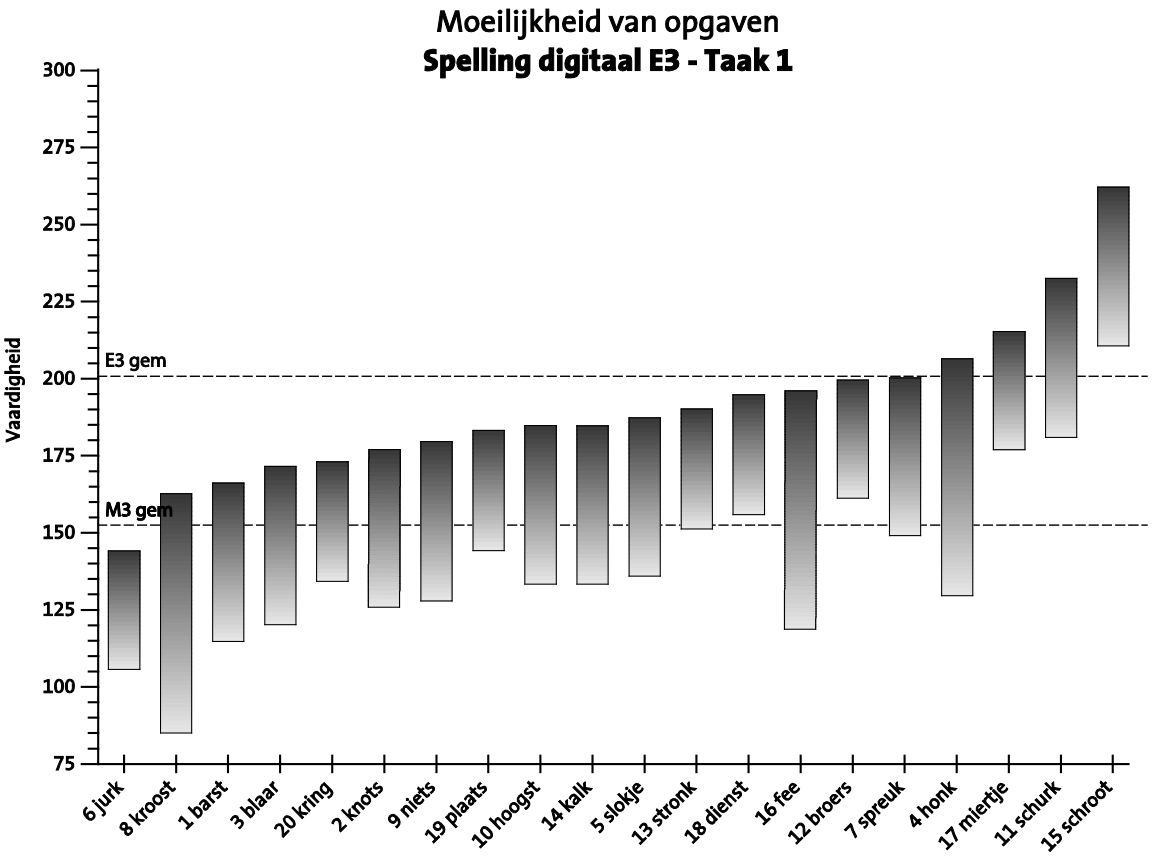
Bijlage 1 Moeilijkheid van opgaven per taak in Spelling 3.0 digitaal groep 3

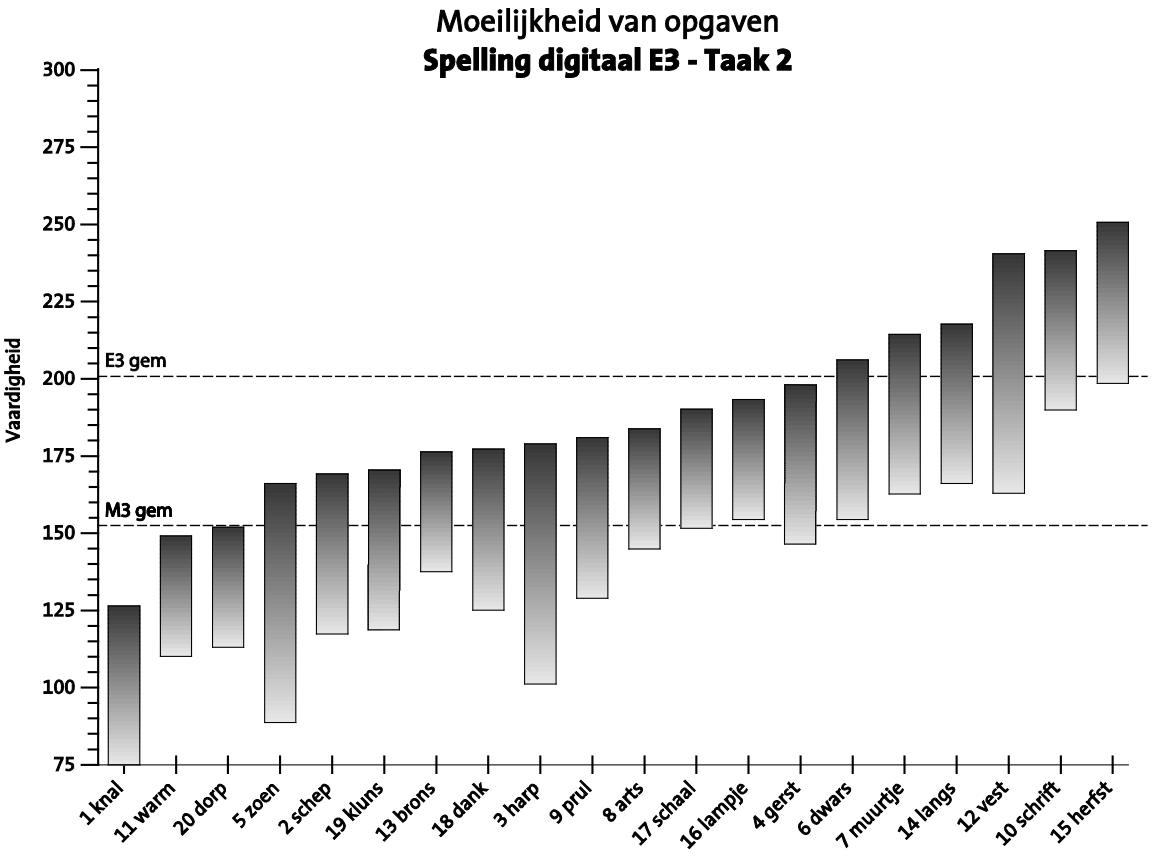





Moeilijkheid van opgaven
Spelling digitaal M3E3 - Taak 1
Vaardigheid
275
250
225
200
175
150
125
100
75
50
E3 gem
M3 gem
1 maat
19 duur
20 jurk
5 bril
6 warm
9 klomp
2 kroost
7 korst
15 paars
11 snoer
14 gang
13 trots
16 korf
8 bank
10 schelp
12 golf
18 dienst
4 wenk
3 spreuk
17 schrik


Moeilijkheid van opgaven
Spelling digitaal M3E3 - Taak 2
Vaardigheid
275
250
225
200
175
150
125
100
75
50
E3 gem
M3 gem
20 pomp
10 plant
1 tulp
6 groot
16 schaap
17 blaar
7 schep
8 straal
4 kunst
2 kring
18 worp
13 grens
12 reus
15 hoogst
9 bonk
5 kreeft
19 triest
11 helft
3 schuit
14 langs


Moeilijkheid van opgaven
Spelling digitaal E3 - Taak 1
Vaardigheid
300
275
250
225
200
175
150
125
100
75
E3 gem
M3 gem
6 jurk
8 kroost
1 barst
3 blaar
20 kring
2 knots
9 niets
19 plaats
10 hoogst
14 kalk
5 slokje
13 stronk
18 dienst
16 fee
12 broers
7 spreuk
4 honk
17 miertje
11 schurk
15 schroot


Moeilijkheid van opgaven
Spelling digitaal E3 - Taak 2
Vaardigheid
300
275
250
225
200
175
150
125
100
75
E3 gem
M3 gem
1 knal
11 warm
20 dorp
5 zoen
2 schep
19 kluns
13 brons
18 dank
3 harp
9 prul
8 arts
17 schaal
16 lampje
4 gerst
6 dwars
7 muurtje
14 langs
12 vest
10 schrift
15 herfst

Bijlage 2 Klassieke en IRT-indices van de opgaven in digitale toetsen Spelling 3.0 groep 3

Toets M3

Volgnr	P-Val	RIT	Beta	Info
1.1	0,927	0,233	-1,325	0,253
1.2	0,697	0,361	-0,375	0,731
1.3	0,917	0,246	-1,247	0,284
1.4	0,742	0,351	-0,502	0,670
1.5	0,760	0,468	-0,397	1,274
1.6	0,591	0,488	-0,053	1,622
1.7	0,825	0,437	-0,567	1,037
1.8	0,794	0,454	-0,482	1,158
1.9	0,737	0,353	-0,489	0,677
1.10	0,777	0,340	-0,610	0,612
1.11	0,801	0,331	-0,693	0,567
1.12	0,852	0,418	-0,648	0,923
1.13	0,917	0,346	-0,911	0,585
1.14	0,845	0,423	-0,628	0,951
1.15	0,774	0,552	-0,356	1,914
1.16	0,636	0,574	-0,103	2,420
1.17	0,556	0,484	0,011	1,652
1.18	0,662	0,574	-0,147	2,352
1.19	0,645	0,488	-0,155	1,548
1.20	0,622	0,489	-0,112	1,584
2.1	0,698	0,679	-0,166	3,900
2.2	0,738	0,475	-0,348	1,338
2.3	0,797	0,605	-0,359	2,469
2.4	0,771	0,617	-0,309	2,652
2.5	0,666	0,366	-0,291	0,767
2.6	0,707	0,678	-0,180	3,846
2.7	0,726	0,566	-0,261	2,131
2.8	0,723	0,630	-0,225	2,937
2.9	0,703	0,633	-0,190	3,043
2.10	0,770	0,617	-0,307	2,658
2.11	0,702	0,571	-0,217	2,224
2.12	0,686	0,485	-0,237	1,469
2.13	0,615	0,573	-0,069	2,464
2.14	0,802	0,602	-0,370	2,428
2.15	0,730	0,566	-0,268	2,117
2.16	0,805	0,601	-0,377	2,403
2.17	0,637	0,677	-0,076	4,204
2.18	0,745	0,625	-0,262	2,813
2.19	0,779	0,550	-0,366	1,890
2.20	0,810	0,534	-0,434	1,720

Toets M3E3

Volgnr	P-Val	RIT	Beta	Info
1.1	0,958	0,234	-0,652	0,338
1.2	0,836	0,261	-0,332	0,512
1.3	0,901	0,221	-0,639	0,341
1.4	0,887	0,333	-0,258	0,800
1.5	0,917	0,302	-0,386	0,620
1.6	0,949	0,391	-0,270	0,976
1.7	0,881	0,423	-0,104	1,353
1.8	0,862	0,247	-0,443	0,446
1.9	0,890	0,414	-0,133	1,271
1.10	0,783	0,389	0,041	1,288
1.11	0,916	0,450	-0,136	1,463
1.12	0,784	0,389	0,038	1,282
1.13	0,868	0,244	-0,470	0,431
1.14	0,928	0,368	-0,271	0,906
1.15	0,864	0,351	-0,180	0,922
1.16	0,841	0,365	-0,110	1,035
1.17	0,921	0,378	-0,242	0,977
1.18	0,739	0,400	0,138	1,441
1.19	0,834	0,454	0,019	1,731
1.20	0,919	0,447	-0,144	1,431
2.1	0,846	0,362	-0,125	1,012
2.2	0,767	0,477	0,157	2,159
2.3	0,932	0,362	-0,287	0,868
2.4	0,884	0,420	-0,113	1,328
2.5	0,860	0,353	-0,167	0,942
2.6	0,866	0,434	-0,063	1,478
2.7	0,755	0,396	0,104	1,389
2.8	0,816	0,377	-0,041	1,152
2.9	0,761	0,478	0,167	2,190
2.10	0,531	0,398	0,512	1,804
2.11	0,863	0,351	-0,178	0,925
2.12	0,721	0,402	0,173	1,494
2.13	0,792	0,386	0,019	1,250
2.14	0,893	0,412	-0,139	1,251
2.15	0,772	0,392	0,066	1,329
2.16	0,854	0,443	-0,029	1,580
2.17	0,877	0,426	-0,093	1,387
2.18	0,792	0,386	0,020	1,253
2.19	0,678	0,406	0,257	1,607
2.20	0,822	0,374	-0,057	1,124

Toets E3

Volgnr	P-Val	RIT	Beta	Info
1.1	0,833	0,368	-0,087	1,074
1.2	0,829	0,370	-0,074	1,095
1.3	0,857	0,352	-0,158	0,957
1.4	0,947	0,250	-0,560	0,421
1.5	0,868	0,244	-0,470	0,431
1.6	0,736	0,489	0,211	2,315
1.7	0,814	0,462	0,063	1,870
1.8	0,731	0,407	0,154	1,465
1.9	0,860	0,350	-0,168	0,941
1.10	0,841	0,446	0,002	1,677
1.11	0,840	0,364	-0,105	1,044
1.12	0,833	0,263	-0,322	0,517
1.13	0,815	0,377	-0,038	1,155
1.14	0,921	0,365	-0,242	0,977
1.15	0,932	0,349	-0,287	0,868
1.16	0,915	0,374	-0,218	1,037
1.17	0,755	0,401	0,104	1,389
1.18	0,765	0,398	0,082	1,354
1.19	0,816	0,377	-0,041	1,152
1.20	0,810	0,463	0,070	1,892
2.1	0,871	0,342	-0,203	0,885
2.2	0,510	0,413	0,547	1,809
2.3	0,761	0,482	0,167	2,190
2.4	0,782	0,475	0,128	2,070
2.5	0,444	0,401	0,657	1,790
2.6	0,557	0,417	0,469	1,789
2.7	0,604	0,419	0,389	1,741
2.8	0,863	0,348	-0,178	0,925
2.9	0,841	0,363	-0,108	1,039
2.10	0,678	0,416	0,257	1,607
2.11	0,854	0,437	-0,029	1,580
2.12	0,759	0,291	-0,073	0,669
2.13	0,784	0,475	0,123	2,058
2.14	0,646	0,311	0,227	0,823
2.15	0,790	0,281	-0,170	0,610
2.16	0,861	0,248	-0,435	0,450
2.17	0,806	0,382	-0,015	1,194
2.18	0,695	0,414	0,226	1,567
2.19	0,769	0,480	0,152	2,145
2.20	0,649	0,499	0,351	2,654

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