1. Uitgangspunten van de toetsconstructie

Bij onderstaande beoordeling van de kwaliteitsaspecten met bijbehorende codes van het voornoemde beoordelingskader worden passages uit de wetenschappelijke verantwoording en de Handleiding veelal letterlijk vermeld. De wetenschappelijke verantwoording heeft betrekking op de uitgangspunten van de toetsconstructie, de normen, de betrouwbaarheid en meetnauwkeurigheid en de validiteit. De Handleiding heeft betrekking op het gebruik van de toets, communicatie over de toetsgegevens en de inhoudsverantwoording.

Algemeen

Het Cito Volgsysteem primair en speciaal onderwijs beoogt de vorderingen van individuele leerlingen, groepen leerlingen en het onderwijs op school van groep 1 tot en met groep 8 te volgen en te evalueren. De digitale toetsen Spelling 3.0 voor groep 4 zijn een onderdeel van het Cito Volgsysteem primair en speciaal onderwijs en zijn bedoeld voor leerlingen in groep 4 van het primair onderwijs. De digitale toetsen Spelling zijn varianten van de papieren toetsen Spelling. Onderstaande beschrijving is gebaseerd op de Handleiding.

Meetpretentie

De toetsen in de toets pakketten Spelling 3.0 van het Cito Volgsysteem primair en speciaal onderwijs zijn bedoeld om vast te stellen hoe goed een leerling kan spellen en hoe de spellingvaardigheid van de leerling zich in de loop van de jaren ontwikkelt.
Het vaststellen van de spellingvaardigheid bij de digitale toetsen Spelling gebeurt door de leerling woorden te laten intypen (actieve spelling). De spellingregels zelf worden niet expliciet bevraagd. De leerling laat indirect zien dat hij of zij de spellingregels beheerst door de gevraagde woorden correct te typen.

Doelgroep

De digitale toetsen Spelling 3.0 groep 4 zijn bedoeld voor leerlingen in groep 4 van het primair en speciaal onderwijs, maar kunnen ook gebruikt worden voor leerlingen uit andere jaargroepen die werken op het niveau van groep 4 en voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften. Voor deze groepen speciale leerlingen zijn geen afzonderlijke normen vastgesteld en de toetsresultaten van deze leerlingen worden geïnterpreteerd met behulp van de gemiddelde vaardigheidsscores voor leerlingen uit het regulier onderwijs.
Hiernaast moeten leerlingen bekend zijn met de letters op het toetsenbord, waarbij het belang ligt bij de kennis dat het toetsenbord hoofdletters bevat in plaats van kleine letters.

Gebruiksdoel en functie

Doel van de toetsen Spelling 3.0 is het in kaart brengen van het vaardigheidsniveau en de ontwikkeling van de leerlingen op het gebied van de spelling, voor groep 4 specifiek het schrijven van eenvoudige klankzuivere en niet-klankzuivere woorden van één of twee lettergrepen. De toetsen maken het mogelijk om:

- De vaardigheid spelling van zowel individuele leerlingen als groepen leerlingen (groeps- en schoolniveau) te beoordelen via een vergelijking van de behaalde scores met de scores van een landelijke referentiegroep oftewel niveaubepaling.
- De ontwikkeling van de vaardigheid spelling van zowel individuele leerlingen als groepen leerlingen (groeps- en schoolniveau) door de leerjaren heen te volgen oftewel progressiebepaling.

Inhoudelijke theoretische inkadering:
De inhoud van de toetsen Spelling 3.0 is gebaseerd op het domein Begrippenlijst en Taalverzorging, en dan het onderdeel Taalverzorging, beschreven in het Referentiekader Taal en Rekenen. De toetsen sluiten aan bij de indeling die is gehanteerd in het Referentiekader Taal. In de publicatie 'Leerstoflijnen begrippenlijst en taalverzorging beschreven' van de SLO, is aangegeven hoe de opbouw van de leerstoflijnen eruit kan zien voor de verschillende groepen. Voor de inhoud van de toetsen zijn deze uitwerkingen bepalend geweest voor de theoretische basis als voor de indeling van het categorieënoverzicht. Verfijning van de categorieën is mede gebaseerd op basis van analyse van methoden (taal en lezen) die veel gebruikt worden in het basisonderwijs.

Inhoud van het toetspakket
Het toetspakket Spelling 3.0 groep 4 bestaat uit de volgende documenten:

- Handleiding, deze bevat informatie over:
  - de afname van de toets (hfdst. 2),
  - nakijken en verwerken van toetsgegevens (hdfst. 3),
  - interpretatie van de toetsresultaten op leerling- en groepsniveaus (hfdst 4),
  - interpretatie van toetsresultaten op schoolniveau (hfdst 5),
  - theoretisch kader en achtergronden van de toets (hfdst 6),
  - communiceren over toetsresultaten met leerling en ouders (hfdst 7),
  - achtergrondinformatie en veelgestelde vragen (hfdst 8) en
  - enkele bijlagen
- Vier toetsen:
  - Toets M4 (Medio groep 4)
  - Toets E4 (Eind groep 4)
  - Toets E3M4 (makkelijke variant van de toets M4)
  - Toets M4E4 (makkelijke variant van de toets E4)
- Nakijkkaarten
- Antwoordbladen
- Tabellen voor de vier toetsen voor het bepalen van de vaardigheidsscore en –niveau.

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor de psychometrische aspecten van (reeksen van) toetsen uit leerlingvolgsystemen (LOVS)', zoals opgesteld door de Expertgroep Toetsen PO. De Expertgroep Toetsen PO wordt gevormd door Prof. Dr. Cees Van der Vleuten (voorzitter), Prof. dr. Cees Glas (psychometrisch expert), Dr. Desiree Joosten-Ten Brinke (onderwijskundig expert) en mevrouw Paulyn K. Berding-Oldersma MSc (secretaris).

De kwaliteit van de steekproef

S1.1. Is de steekproef representatief?

Bevindingen:
Sinds 2013/2014 wordt door Cito een nieuwe werkwijze voor het normeren van leerlingvolgsystemen toegepast. Een belangrijk aspect van die werkwijze is om de normeringssteekproef (voor 50 procent) te baseren op gegevens uit het zogenaamde embedded field onderzoek en (voor 50 procent) op gegevens uit Cito dataretour. Dataretour is de optie waarbij scholen Cito toestemming geven om de toetsgegevens van leerlingen te gebruiken voor toetsconstructie-, normerings- en onderzoeksdoeleinden. Bij de normering van de derde generatie LVS-toetsen is rekening gehouden met de variabelen regio, urbanisatiegraad, schooltype en sekse. Op pagina 46-47 van de verantwoording Spelling 3.0 (papieren versie) voor groep 4 wordt een algoritme beschreven dat een representatieve steekproef moet garanderen. Niettemin werd er naast die garantie een controle op de representativiteit uitgevoerd door de populatieverdelingen uit gegevens van DUO te vergelijken met de steekproefverdelingen. De controle bevestigde dat de normeringssteekproeven een zeer goede afspiegeling van de populatie zijn met effectgroottes ver onder de 0.10. Voor de deelnemende scholen aan het normeringsonderzoek is nagegaan of zij als groep afweken van wat men voor de totale populatie van scholen zou mogen verwachten. Niet alleen de gemiddelde score op de Cito Eindtoets Basisonderwijs, maar ook specifiek de score voor spelling bleek voor de scholen in het normeringsonderzoek niet af te wijken van het populatiegemiddelde voor de Eindtoets. Dit is voor meerdere jaren onderzocht.
In tabel 4.7 van de verantwoording Spelling 3.0 (digitale versie) voor groep 4 worden de uiteindelijke aantallen leerlingen gegeven die meegenomen zijn bij de normering. Voor M4 betreft dat 6717 leerlingen (3115 leerlingen uit de steekproef en 3602 leerlingen via dataretour) afkomstig van 264 scholen en voor E4 6094 leerlingen (2854 leerlingen uit de steekproef en 3240 leerlingen via dataretour) afkomstig van 220 scholen.

Conclusie:
De representativiteit van de steekproef is onderzocht met betrekking tot regio, urbanisatiegraad, schooltype en sekse. De conclusie is dat de normeringssteekproeven een zeer goede afspiegeling vormen van de populatie. Op aspect S1.1 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

S1.2. In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?

Bevindingen:
Aan het begin van het toetsontwikkelingsproces van de LVS-toetsen Spelling 3.0 groep 4, zijn in 2011 opgaven geconstrueerd. Deze opgaven zijn allereerst in *papieren* versie onderzocht in kalibratieonderzoeken 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 4 paragraaf 4.2 van de wetenschappelijke verantwoording van de LVS-toetsen Spelling (papieren versie) voor groep 4 (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.

In januari 2014 vond het papier-digitaal-kalibratieonderzoek M4 plaats en in juni 2014 vond het papier-digitaal-kalibratieonderzoek E4 plaats. In deze onderzoeken werd slechts een deel van de items meegenomen die geselecteerd waren voor de papieren uitgaven Spelling 3.0 groep 4, omdat het onderzoekstechnisch niet mogelijk was om alle papieren items mee te nemen. Om alle items te kunnen onderzoeken, zouden namelijk grote aantallen scholen en leerlingen nodig zijn geweest. Op basis van ervaringen met het werven van scholen voor andere proeftoetsen die digitaal werden afgenomen, was bekend dat de bereidheid onder scholen om deel te nemen aan dergelijke digitale onderzoeken zeer laag was. Noodgedwongen is om die reden gekozen voor een opzet en design waarbij zo min mogelijk leerlingen nodig waren, maar waarbij nog wel toetsen van goede kwaliteit samengesteld konden worden. Alleen de opgaven van de reguliere papieren toetsen 3.0 (d.w.z. de toetsen die passen bij de reguliere afnamemomenten M4 en E4, dus niet de tussentoetsen E3M4 en M4E4) zijn omgezet naar digitale versies. De reguliere papieren tussentoetsen E3M4 en M4E4 zijn slechts gedeeltelijk omgezet naar digitale versies. Voor deze uitgaven zijn vervolgens naast nieuwe opgaven ook opgaven uit de digitale Starttaak LVS Spelling tweede generatie, die eveneens in de kalibratieonderzoeken werd meegenomen, geselecteerd.

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 van de wetenschappelijke verantwoording van de digitale versie) 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. De vaardigheid op het papieren gedeelte van de toets kan vergelijkbaar worden geacht met de vaardigheid op het digitale gedeelte van de toets, omdat de spellingvaardigheid van een leerling niet verandert tijdens een toets. Hierdoor zijn de twee afnamemethoden verbonden. Wel zijn eigen moeilijkheids-en discriminatieparameters geschat (onder het IRT-model OPLM) voor de digitale items. Immers, 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. De papieren versie en digitale versie van een item hoeven immers niet noodzakelijkerwijs precies even moeilijk te zijn en/of even goed te discrimineren.

In het papier-digitaal onderzoek voor het ‘medio’ afnamemoment (M4) van januari 2014 zijn 110 items voorgelegd aan 702 leerlingen van groep 4. 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 M4 3.0. Daarnaast maakte elke leerling de Starttaak M4 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 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 de papieren en digitale items worden vergeleken: passen ze op een en dezelfde schaal? Doordat ook een grote groep leerlingen beide taken digitaal maakte, was

men 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 via het normeringsonderzoek van de papieren toets 'verbonden design. De data van het papier-digitaal onderzoek werden gekoppeld aan de data die verzameld werden in het normeringsonderzoek voor de papieren uitgave M4 uit 2013. Om de koppeling te realiseren werden de data die verzameld waren 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).

In het papier-digitaal onderzoek voor het 'einde' afnamemoment (E4) van juni 2014 zijn 110 items voorgelegd aan 651 leerlingen van groep 4 (zie Tabel 4.2 van de wetenschappelijke verantwoording van de digitale versie). 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 E4 3.0. Daarnaast maakte elke leerling de Starttaak E4 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 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 werd voor gezorgd dat geen enkele leerling beide versies (papier en digitaal) van een eenzelfde item kreeg voorgelegd. Doordat de helft van de onderzoeksgroep de Starttaak van de tweede generatie op papier maakte, konden de papieren en digitale items worden vergeleken: passen ze op een en dezelfde schaal? Doordat ook een grote groep leerlingen beide taken digitaal maakte, was men 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 E4 uit 2013. Om de koppeling te realiseren werden de data die verzameld waren in het papier-digitaal onderzoek toegevoegd aan de dataset van het genormeerde normeringsonderzoek (die dient voor de schaling van de items in de itembank Spelling).

Conclusie:
Het onvolledige dataverzamelingsdesign is adequaat. 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 en dat er dus sprake is van een eendimensionale vaardigheidsschaal waar items en leerlingen op afgebeeld kunnen worden. Met andere woorden, zowel de items als de gehele toets van de digitale versie passen bij het IRT-model OPLM (was al aangetoond voor de papieren versie). Ook kan geconcludeerd worden dat de papieren en digitale versies van toetsen van hetzelfde niveau (bijvoorbeeld M4) leiden tot dezelfde vaardigheidsschatting en dat dezelfde normering gehandhaafd kan worden (deze lag al vast voor de papieren toetsen) en er dus geen onderzoek nodig is naar de equivalentie van de scores op de verschillende versies: de scores zijn immers per definitie uitwisselbaar. Op aspect S1.2 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

Normering

N1.2.1. Zijn de normgroepen groot genoeg?

Bevindingen:

Uit de afgenomen taken werden de digitale toetsen E3M4, M4, M4E4 en E4 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 item functioneren (DIF) ten opzichte van de papieren items. Daarom was er onvoldoende vertrouwen in het zonder meer kunnen toepassen van de 'papieren normering' op de digitale toetsversies en 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 in 2016. Tabel 4.3 van de wetenschappelijke verantwoording van de digitale versie laat zien dat de aantallen via dataretour gerealiseerde afnames sterk verschilden. Met name de tussentoetsen E3M4 en M4E4 werden relatief weinig afgenomen, wat gezien de functie van deze toetsen begrijpelijk is. De totale aantallen afnames die in de analyses konden worden meegenomen volstonden echter voor een nauwkeurige kalibratie, nl. 290 voor E3M4, 2039 voor M4, 249 voor M4E4 en 1904 voor E4 (zie de laatste kolom van Tabel 4.3 van de wetenschappelijke verantwoording van de digitale versie).
In tabel 4.8 van de wetenschappelijke verantwoording van de digitale versie wordt voor M4 en E4 de normtabel met relatieve normen op leerlingniveau gepresenteerd voor Spelling 3.0 groep 4. Voor aantallen leerlingen en scholen zie S1.1. Voor beide afnamemomenten worden vaardigheidsverdelingen gepresenteerd, d.w.z. gemiddelde score, standaarddeviatie, kurtosis, scheefheid en de percentielen P10, P20, P25, P40, P50, P60, P75, P80. Met behulp van deze percentielen kunnen de twee niveau-indelingen (A t/m E en I t/m V) opgesteld worden.

Conclusie:
De normgroepen zijn groot genoeg (zie S1.1). Op aspect N1.2.1 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

N1.2.2. Zijn de normgroepen representatief?

Bevindingen:
De representativiteit van de steekproeven werd in S1.1 en S2.1 besproken en daar werd geconstateerd dat de steekproeven representatief waren voor regio, urbanisatiegraad, schooltype en sekse.
Voor Spelling 3.0 (digitaal) groep 4 geldt dat de normen geldig zijn tot en met 2023.

Conclusie:
Op aspect N1.2.1 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

Betrouwbaarheid

B1.1. Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
In hoofdstuk 5 van de wetenschappelijke verantwoording van de digitale versie wordt beschreven hoe, gebruikmakend van het feit dat de items uit de toets geschaald zijn met OPLM, een betrouwbaarheidscoëfficiënt, de MAcc (Accuracy of Measurement), berekend kan worden die qua interpretatie grote overeenkomst vertoont met de

betrouwbaarheidscoëfficiënt uit de klassieke testtheorie (KTT). Deze coëfficiënt wordt in de psychometrische literatuur beschreven en als correct aangemerkt.

Conclusie:
De betrouwbaarheidsgegevens worden correct berekend. Op aspect B1.1 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

B1.2. Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets genomen worden?

Bevindingen:
Voor E3M4, M4, M4E4 en E4 zijn drie betrouwbaarheidsgegevens berekend: standaardmeetfout, MAcc en een gesimuleerde test-hertest betrouwbaarheidscoëfficiënt. Voor E3M4 zijn de gegevens gelijk aan 8,74, 0,91 en 0,91, voor M4 gelijk aan 9,95, 0,92 en 0,92, voor M4E4 gelijk aan 9,96, 0,92 en 0,92 en voor E4 gelijk aan 11,24, 0,93 en 0,93. De auteurs van de wetenschappelijke verantwoording van de digitale versie verwijzen naar het beoordelingssysteem van de COTAN waar voor tests die geen zware consequenties voor leerlingen hebben, zoals de toetsen Spelling 3.0 (digitaal) groep 4, een betrouwbaarheidscoëfficiënt van meer dan 0,80 als 'goed' aangemerkt wordt. Net als bij de papieren toetsen, zijn de betrouwbaarheidscoëfficiënten (MAcc's en test-hertest) voor de digitale versie van de toetsen voor groep 4 zeer hoog. Ze variëren van 0,91 tot 0,93.
Naast klassieke betrouwbaarheidscoëfficiënten is ook de lokale betrouwbaarheid en de meetnauwkeurigheid onderzocht. De betekenis van de meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit de betrouwbaarheidstabellen van twee niveauverdelingen (I t/m V en A t/m E) voor E3M4, M4, M4E4 en E4. De informatie uit de betrouwbaarheidstabellen worden samengevat in twee indices: de plus/minus 1 niveau-index en de marginal classification accuracy. Uit Tabel 5.6 van de wetenschappelijke verantwoording van de digitale versie blijkt dat gemiddeld gezien, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 97,6 tot 99,9 procent van de leerlingen in een scoregroep ook in werkelijkheid scoort in die scoregroep, ***of*** één scoregroep daarboven ***of*** één scoregroep daaronder. Uit deze indices blijkt dat gegeven het gebruiksdoel van de toetsen de classificaties voldoende betrouwbaar zijn, d.w.z. dat gebruikers rekening dienen te houden met maximaal 1 niveau verschil en dus is het percentage misclassificaties erg beperkt. De *Marginal Classification Accuracy* loopt uiteen van 63,3 tot 75,3 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij beide afnamemomenten gemiddeld gezien in zo'n 69 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. Daarnaast werd inzicht gegeven in de lokale betrouwbaarheid: de meetfout blijkt het kleinst te zijn in de lagere en gemiddelde vaardigheidsregionen.

Conclusie:
De betrouwbaarheid van de toetsen Spelling 3.0 (digitaal) groep 4 is 'voldoende' (en dus kunnen de leerlingen op een betrouwbare manier ingedeeld worden in normgroepen) als aangenomen mag worden dat de toets geen zware consequenties voor de leerlingen heeft en ingestemd wordt met de beoordelingscriteria voor de betrouwbaarheid van de COTAN.

Op basis van het voorgaande wordt op aspect B1.2 aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

Validiteit

V1. Dragen de items in de toets bij aan de validiteit van de toets (hierbij gaat het om aspecten als relevantie, objectiviteit en efficiëntie van de items)

Bevindingen:
In de groepen 3 en 4, bij de start van het formele lees- en spellingonderwijs, zijn de toetsen bedoeld om vast te stellen hoe goed een leerling eenvoudige klankzuivere en niet-klankzuivere woorden van één (gr 3 en 4) of twee (gr 4) lettergrepen kan schrijven. Dit gebeurt door actieve toetsing (woord- en zinsdictee) wat een goede keuze is. Een verschil met klassikale toetsing waarbij de leerkracht voorleest, is dat leerlingen in de digitale versie zelf kunnen kiezen om een item te laten herhalen.

Logischerwijs wordt een beperkt aantal spellingcategorieën getoetst. De zinnen in de toetsen zijn ook gezien de doelgroep kort (bijv. Onze hond is heel trouw) en de woorden die de kinderen opschrijven zijn voor de meeste kinderen bekend en passen binnen de spellingcategorieën die worden gehanteerd. In de handleiding wordt via tabellen inzichtelijk gemaakt welke spellingcategorieën worden gehanteerd en welke woorden in de toetsen eronder vallen.

Ook wordt inzicht gegeven in de moeilijkheid van opgaven op basis van vaardigheidsscores. Leervorderingen worden systematisch weergegeven via allerlei rapportages en er worden handvatten geboden voor analyse.

Er wordt een duidelijke relatie gelegd met het Referentiekader door, voor de keuzes in te toetsen spellingcategorieën, gebruik te maken van de uitwerking in leerstoflijnen die door de SLO ontwikkeld werden om de weg naar 1F en 2F te beschrijven. In de toetsen die nog op de markt komen voor de bovenbouw zal echter duidelijk moeten worden of de school kan nagaan of een leerling referentieniveau 1F of 1S behaalt.

Voor de verantwoording van de validiteit wordt verwezen naar hoofdstuk 6 van de Wetenschappelijke verantwoording van de LVS-toetsen Spelling groep 4 (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 van de Wetenschappelijke verantwoording van de digitale versie) en dat daarmee net als bij de papieren versie voldaan wordt aan eisen van unidimensionaliteit als waarborg voor de constructvaliditeit van de toetsen.

Conclusie:
Op aspect V1.1 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 op dit aspect het oordeel '**voldoende**' toegekend.

Het volg-aspect

VA1.1. Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een adequate manier gemeten?

Bevindingen:
De resultaten van het kalibratieonderzoek (zie ook S1.2) laten zien dat de items van de verschillende toetsen Spelling 3.0 groep 4 op een eendimensionale vaardigheidsschaal afgebeeld kunnen worden en dat aan de hand van de door de leerling behaalde vaardigheidsscores op de onderscheiden toetsen diens groei adequaat gemeten kan worden.

Conclusie:
Op aspect VA1.1 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

VA1.2. Worden er gegevens verstrekt over hoe groei geïnterpreteerd dient te worden? Wordt de betrouwbaarheid van de groei op die schaal adequaat weergegeven?

Bevindingen:
In de Handleiding wordt aan de hand van een leerlingrapport de interpretatie van groei op een duidelijke manier beschreven. De 67 procent betrouwbaarheidsintervallen van de vaardigheidsscores worden zowel op het leerlingrapport als in afzonderlijke tabellen vermeld. Groei wordt zowel met vergelijkingen tussen verschillende leerlingen als met vergelijkingen op andere tijdstippen van dezelfde leerlingen weergegeven. Zowel in de wetenschappelijke verantwoording en de handleiding wordt het volgen van groei van leerlingen adequaat toegelicht.

Conclusie:
Op aspect VA1.2 wordt aan de toetsen Spelling 3.0 (digitaal) groep 4 het oordeel '**voldoende**' toegekend.

Inzicht in leervorderingen

I1. Levert de toetsaanbieder een format voor een geschreven toelichting bij de leervorderingen van de leerling die (ook) voor ouders/voogden/verzorgers begrijpelijk is?

Via de portal van Cito B.V. kan gebruik worden gemaakt van rapportage-/registratieformulieren voor een leerlingrapport, groepsrapport, groepsoverzicht (overzicht van één groep leerlingen tijdens hun schoolperiode) en een alternatief leerlingrapport (voor leerlingen die op een eigen niveau werken).
In de Handleiding wordt in hoofdstuk 7 aandacht besteed aan hoe met ouders over de toetsresultaten gecommuniceerd kan/moet worden. Met name wordt daarbij gewezen op het leerlingrapport waarin zowel het niveau van de leerling als de progressie van de leerling numeriek en grafisch gepresenteerd worden. Hiervoor is ook een folder ouderinformatie beschikbaar die men via de website van het Cito kan downloaden.

Daarnaast wordt de docent gewezen op misverstanden die zich bij de interpretatie van de niveau-indelingen bij de ouders kunnen voordoen. Ook moeten zij aan ouders het verschil tussen methode-onafhankelijke en methodegebonden toetsen duidelijk maken en erop wijzen dat deze toetsen leerlingen anders (kunnen) beoordelen. In hoofdstuk 8 worden ook veelgestelde vragen behandeld die weliswaar voor de docenten bestemd zijn maar zij kunnen met die informatie bijvoorbeeld via tienminutengesprekken ook de ouders beter voorlichten.
De wijze waarop de registratieformulieren zijn vormgegeven en de uitleg die voor docenten beschikbaar is geven het niveau en de groei van de individuele leerling weer.

Conclusie:
Op aspect I1.1 wordt aan de toetsen Spelling 3.0 groep 4 het oordeel '**voldoende**' toegekend.

3. Verzamelstaat

Kwaliteitsaspect	**Code**	**Oordeel**
De kwaliteit van de steekproef	*S1.1*	**Voldoende**
	S1.2	**Voldoende**
Normering	*N1.1*	**Voldoende**
	N1.2	**Voldoende**
Betrouwbaarheid	*B1.1*	**Voldoende**
	B1.2	**Voldoende**
Validiteit	*V1*	**Voldoende**
Volg-aspect	*VA1.1*	**Voldoende**
	VA1.2	**Voldoende**
Inzicht in leervorderingen	*I1*	**Voldoende**

4. Literatuurlijst

Bij deze beoordeling zijn de volgende, door Cito B.V. verstrekte, materialen gebruikt:

- M. Tomesen, J. Wouda, A. Mols & L. Horsels (2015). *Wetenschappelijke verantwoording van de LVS-toetsen Spelling 3.0 voor groep 4*. Arnhem: Cito B.V.
- M. Tomesen, J. Wouda, L. Horsels (2017). *Wetenschappelijke verantwoording Spelling 3.0 digitaal voor groep 4.* Arnhem: Cito B.V.
- *Leerkrachtmap Spelling 3.0 voor groep 4.* Arnhem: Cito B.V.