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 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 toetsen Taalverzorging groep 6 tot en met 8 zijn een onderdeel van het Cito Volgsysteem primair onderwijs en zijn bedoeld voor leerlingen in groep 6, 7 en 8 van het primair onderwijs. Onderstaande beschrijving is gebaseerd op de Handleiding en de Wetenschappelijke verantwoording.

Meetpretentie

De toetsen in het toets pakket Taalverzorging groep 6 tot en met 8 van het Cito Volgsysteem primair en speciaal onderwijs zijn bedoeld om vast te stellen hoe goed een leerling de juiste spelling- en interpunctieregels kan toepassen en hoe de grammaticale kennis van de leerling zich in de loop van de jaren ontwikkelt.
Het vaststellen van de vaardigheid in de verschillende taalverzorgingscomponenten gebeurt door de leerling beslissingen te laten nemen over getoonde spelling- en interpunctiemogelijkheden. De spellingregels zelf worden niet expliciet bevraagd. De leerling laat indirect zien dat hij of zij de spellingregels beheerst door de correct geschreven woorden (spelling) en zinnen (interpunctie) te herkennen.

Doelgroep

De toetsen Taalverzorging groep 6 tot en met 8 zijn bedoeld en genormeerd voor leerlingen in groep 6, 7 en 8 van het primair onderwijs, maar kunnen ook gebruikt worden voor leerlingen uit andere jaargroepen die werken op het niveau van groep 6, 7 en 8 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.

Voor leerlingen die nog maar pas in Nederland verblijven, zijn de toetsen ongeschikt. Een leerling moet voldoende taalvaardig in het Nederlands zijn (vergelijkbaar met Nederlandstalige leerlingen in groep 6, 7 en 8) voordat de toets Taalverzorging kan worden afgenomen.

Gebruiksdoel en functie

Doel van de toetsen Taalverzorging groep 6 tot en met 8 is het in kaart brengen van het vaardigheidsniveau en de ontwikkeling van de leerlingen op het gebied van geschreven taal verzorging, voor groep 6, 7 en 8. De toetsen maken het mogelijk om:

- De vaardigheid in de verschillende taalverzorgingscomponenten 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 in de verschillende taalverzorgingscomponenten 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 Taalverzorging groep 6 tot en met 8 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. Taalverzorging wordt onderverdeeld in vier deelgebieden, te weten spelling niet-werkwoorden, spelling werkwoorden, interpunctie en grammatica. Toetsing van de spelling van werkwoorden wordt in de toetsen Taalverzorging groep 7 en 8 gedaan, aangezien werkwoordspelling nog niet voldoende onderwezen wordt in groep 6. De koppeling die gemaakt wordt met het Referentiekader levert voor scholen een meerwaarde. Het rapporteren van de referentieniveaus verschaft inzicht over wat de niveaus 1F en 2F voor taalverzorging inhouden en hoe leerlingen zich met hun toetsresultaten verhouden tot de referentieniveaus.

Inhoud van het toetspakket
Het toetspakket Taalverzorging groep 6 tot en met 8 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
- Drie toetsen:
  - Toets M6/E6
  - Toets M7/E7
  - Toets M8
- Afnamekaarten met aanwijzingen voor de papieren of de digitale afname van de toetsen
- Nakijkkaarten
- Antwoordbladen
- Tabellen voor de drie 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. Desirée Joosten-Ten Brinke (onderwijskundig expert) en mevrouw Paulyn Berding-Oldersma MSc (secretaris).

De kwaliteit van de steekproef

S1.1. Is de steekproef representatief?

Bevindingen:
Bij proeftoetsingen is in 2013 halverwege de leerjaren 6, 7 en 8 een aantal opgaven voorgelegd aan leerlingen. Het proeftoetsdesign was een onvolledig design maar het was wel verbonden door middel van blokjes met ankeropgaven, d.w.z. niet alle leerlingen in de steekproef van het kalibratieonderzoek maakten alle opgaven. Op het afnamemoment M6 zijn 211 opgaven geproeftoetst verdeeld over 7 boekjes (booklets). Op het afnamemoment M7 zijn 280 nieuwe opgaven verdeeld over 7 boekjes (booklets) en op het afnamemoment M8 272 nieuwe opgaven verdeeld over 6 boekjes (booklets). Elke deelnemende school maakte meerdere taken. Een boekje (booklet) voor leerjaar 6 bestond uit drie taken: een taak spelling niet-werkwoorden (20 opgaven), een taak interpunctie (20 opgaven) en een taak grammatica (20 opgaven). Een boekje (booklet) voor leerjaar 7 bestond naast deze taken ook nog uit een taak spelling werkwoorden (20 opgaven). Een boekje (booklet) voor leerjaar 8 bestond uit dezelfde taken als een boekje (booklet) voor leerjaar 7, met als verschil dat de taak grammatica uit 30 opgaven bestond in plaats van 20. De nieuwe opgaven werden door minimaal 280 leerlingen maar het merendeel van de opgaven is door minimaal 600 leerlingen gemaakt. Na de afnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programmapakket One Parameter Logistic Model (OPLM).

Op grond van het uitgevoerde kalibratieonderzoek in 2013, zijn in januari/begin februari 2014 en 2015 (M-momenten) en mei/begin juni 2014 en 2015 (E-momenten), items geselecteerd voor drie toetsen, te weten de toets voor leerjaar 6 (toets M6/E6), leerjaar 7 (toets M7/E7) en leerjaar 8 (toets M8). De voor deze toetsen bedoelde items werden samen met een aantal aanvullende items in 2014 en 2015 voorgelegd aan in totaal ruim 6200 leerlingen (op M6 799 leerlingen, op M7 1812, op M8 941 en op E6 en E7 respectievelijk 1323 en 1346 leerlingen), verdeeld over in totaal 222 scholen (op M6 31 scholen, op E6 50, op M7 56, op E7 50 en op M8 35 scholen). Het ging in totaal (dus inclusief de aanvullende opgaven) om 418 items. De opgaven in het normeringsonderzoek werden op dusdanige wijze aan de leerlingen voorgelegd dat het mogelijk was in de kalibraties alle items over de afnamemomenten heen met elkaar te verbinden. Ter illustratie van dit onvolledige, maar geheel verbonden design, halen we nu het normeringsonderzoek M7 van januari 2015 even naar voren. Er zijn in totaal 155 items opgenomen in het onderzoek. Deze items waren verdeeld over vijf verschillende opgavenboekjes in een onvolledig, maar 'verbonden' design. Elk boekje bestond uit 80 items, 20 items per deelgebied. Elk item kwam in principe in twee boekjes voor. Soms kwam een blokje maar een keer voor in het design van deze jaargroep, maar dat blokje werd dan wel weer opgenomen in het onderzoek van een aangrenzend afnamemoment (E6 of E7). Het gemiddeld aantal leerlingantwoorden per item was 426.

De genoemde aantallen waren ruimschoots voldoende voor het doel van het onderzoek (kalibreren en normeren). Hiervoor was een minimum van 400 waarnemingen per item vereist; voor het merendeel van de opgaven werd aan deze minimumeis voldaan. Bij een gering aantal items waren er minder dan 400 waarnemingen, hetgeen voor elk onderdeel het geval was bij 10 items. Doordat een gedeelte van de items echter op meerdere tijdstippen werd afgenomen, werd toch aan de minimumeis voldaan.

De representativiteit voor de normeringsonderzoeken M6/E6, M7/E7 en M8 is onderzocht met betrekking tot regio, urbanisatiegraad (verstedelijking), schooltype (strata), schoolgrootte en sekse. Bij regio is uitgegaan van de vier landsdelen / regio's van de CBS-indeling. Bij urbanisatiegraad is uitgegaan van de CBS-indeling naar vijf niveaus van verstedelijking. Bij schooltype is uitgegaan van de formatiegewichten volgens OCW. Hierin worden drie niveaus onderscheiden die gebaseerd zijn op het opleidingsniveau van de ouders. Bij schoolgrootte is een onderscheid gemaakt tussen grote scholen (> 200 leerlingen) en kleine scholen (≤ 200 leerlingen). Bij sekse is een tweedeling gemaakt naar jongens en meisjes.

De steekproefverdeling voor M6/E6, M7/E7 en M8 vormt een zeer goede afspiegeling van de populatieverdeling. De deelnemende scholen zijn dus voor alle achtergrondvariabelen redelijk representatief te noemen voor de populatie van scholen. De effectgroottes liggen allemaal onder de 0.10 en zijn daarmee te interpreteren als klein tot middelgroot. Statistische weging is om die reden dan ook niet nodig. Hetzelfde geldt op leerlingniveau voor de verdeling naar sekse. Schoolgrootte wordt niet beschouwd als een – voor de representativiteit – relevante achtergrondvariabele. Het wordt gebruikt om een onevenwichtige verdeling op andere achtergrondkenmerken zoals regio en urbanisatiegraad te voorkomen.

Conclusie:
De steekproeven zijn representatief, zijn adequaat gestratificeerd naar sekse, regio, urbanisatiegraad en schooltype en geven informatie over hoe de steekproeven zich verhouden tot de populatiewaarden. De procedure voor het samenstellen van de steekproeven is onderbouwd en de omstandigheden waaronder data is verzameld, is redelijk vergelijkbaar met de omstandigheden waaronder de toets wordt afgenomen. Daarmee wordt aan aspect S1.1 het oordeel '**voldoende**' toegekend.

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

Bevindingen:
Om te komen tot een set van psychometrisch en inhoudelijk geschikte items zijn de opgaven uit de proefonderzoeken van januari 2013 (M6, M7 en M8) en de opgaven uit de daaropvolgende normeringsonderzoeken gekalibreerd. Hiervoor is gebruikgemaakt van het IRT model OPLM. Met dit statistische model zijn de psychometrische kenmerken (moeilijkheidsparameters en discriminatie-indices) van de items geschat. In het kalibratieproces is uitgegaan van een onvolledig maar 'verbonden' design.

In het kalibratieproces M6 van januari 2013 zijn 211 nieuwe items voorgelegd aan leerlingen van groep 6. De 211 items waren verdeeld over 7 boekjes (booklets). Elke deelnemende school maakte op het medio-moment meerdere taken. Een boekje voor leerjaar 6 bestond uit drie taken: een taak spelling niet-werkwoorden (20 opgaven), een taak interpunctie (20 opgaven) en een taak grammatica (20 opgaven). De nieuwe opgaven werden door minimaal 280 leerlingen maar het merendeel van de opgaven is door minimaal 600 leerlingen gemaakt, hetgeen boven het minimum vereiste van 400 ligt.

In het kalibratieproces M7 van januari 2013 zijn 280 nieuwe items voorgelegd aan leerlingen van groep 7. De 280 items waren verdeeld over 7 boekjes (booklets). Elke

deelnemende school maakte op het medio-moment meerdere taken. Een boekje voor leerjaar 7 bestond uit vier taken: een taak spelling niet-werkwoorden (20 opgaven), een taak interpunctie (20 opgaven), een taak grammatica (20 opgaven) en een taak spelling werkwoorden (20 opgaven). De nieuwe opgaven werden door minimaal 280 leerlingen maar het merendeel van de opgaven is door minimaal 600 leerlingen gemaakt, hetgeen boven het minimum vereiste van 400 ligt.

In het kalibratieproces M8 van januari 2013 zijn 272 nieuwe items voorgelegd aan leerlingen van groep 8. De 272 items waren verdeeld over 6 boekjes (booklets). Elke deelnemende school maakte op het medio-moment meerdere taken. Een boekje voor leerjaar 8 bestond uit vier taken: een taak spelling niet-werkwoorden (20 opgaven), een taak interpunctie (20 opgaven), een taak grammatica (30 opgaven) en een taak spelling werkwoorden (20 opgaven). De nieuwe opgaven werden door minimaal 280 leerlingen maar het merendeel van de opgaven is door minimaal 600 leerlingen gemaakt, hetgeen boven het minimum vereiste van 400 ligt.

Op basis van inhoudelijke en psychometrische criteria werden 80 items voor elk van de toetsen Taalverzorging groep 6 tot en met 8 geselecteerd (M6/E6, M7/E7 en M8). Voor alle taalverzorgingsonderdelen, te weten interpunctie (IP), spelling werkwoorden (SW), spelling niet-werkwoorden (SN) en grammatica (GR), werden 20 items geselecteerd. Van alle opgaven die zijn meegegaan in het normeringsonderzoek zijn de gekalibreerde p-waarde en rit-waarde bepaald, welke allemaal voldeden aan de minimale vereisten. Voor de normeringsonderzoeken M6/E6, M7/E7 en M8 werden na het trekken van een representatieve steekproef, waarbij rekening werd gehouden met verdeling naar regio, urbanisatiegraad, schooltype en sekse, scholen geworven.

Voor het normeringsonderzoek M6 werd gebruikgemaakt van resultaten van 799 leerlingen uit groep 6 van 31 scholen. Voor het normeringsonderzoek E6 werd gebruikgemaakt van resultaten van 1323 leerlingen uit groep 6 van 50 scholen. Voor het normeringsonderzoek M7 werd gebruikgemaakt van resultaten van 1812 leerlingen uit groep 7 van 56 scholen. Voor het normeringsonderzoek E7 werd gebruikgemaakt van resultaten van 1346 leerlingen uit groep 7 van 50 scholen. Voor het normeringsonderzoek M8 werd gebruikgemaakt van resultaten van 941 leerlingen uit groep 8 van 35 scholen.

Uit het kalibratieonderzoek blijkt dat de items passen bij voornoemd IRT model en dat het model ook past voor de toetsen als geheel. Dit betekent dat er sprake is van één unidimensionele vaardigheidsschalen waar items en leerlingen op afgebeeld kunnen worden.

Conclusie:
Het onvolledige maar 'verbonden' design van de proefonderzoeken is adequaat. Het volledige design van de toetsen M6/E7, M7/E7 en M8 zijn eveneens adequaat. Aan aspect S1.2 wordt het oordeel '**voldoende**' toegekend.

Normering

N1.2.1. Zijn de normgroepen groot genoeg?

Bevindingen:
De toets is genormeerd voor de afnamemomenten M6/E6, M7/E7 en M8.

Op grond van het uitgevoerde kalibratieonderzoek, zijn items geselecteerd voor drie toetsen, te weten de toets voor leerjaar 6 (toets M6/E6), leerjaar 7 (toets M7/E7) en leerjaar 8 (toets M8). De voor deze toetsen bedoelde items werden voorgelegd aan in totaal ruim 6200 leerlingen (op M6 799 leerlingen, op M7 1812, op M8 941 en op E6 en E7 respectievelijk 1323 en 1346 leerlingen), verdeeld over in totaal 222 scholen (op M6 31 scholen, op E6 50, op M7 56, op E7 50 en op M8 35 scholen). Het ging in totaal om 418 items.

Voor de afnamemomenten M6/E6, M7/E7 en M8 werden vaardigheidsverdelingen gepresenteerd op leerlingniveau en op schoolniveau. Dit betreft de gemiddelde score, standaarddeviatie en de percentielen P10, P20, P25, P40, P50, P60, P75, P80 en P90. Van hieruit kunnen de beide niveau indelingen (de symmetrische niveau indeling I t/m V en de asymmetrische niveau indeling A t/m E) worden bepaald.

De normen voor de toetsen 'Taalverzorging groep 6 tot en met 8' zijn geldig tot en met 2024.

Conclusie:
Er is sprake van relatieve normen, de steekproeven zijn representatief en groot genoeg. Daarmee wordt aan aspect N1.2.1. het oordeel '**voldoende**' toegekend.

N1.2.2. Zijn de normgroepen representatief?

Bevindingen:
De representativiteit van de steekproeven is besproken bij punt S1.1. Hier wordt reeds geconstateerd dat deze representatief zijn.

Conclusie:
Aan aspect N1.2.2. wordt het oordeel '**voldoende**' toegekend.

Betrouwbaarheid

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

Bevindingen:
Om relevante gegevens bij de toets te genereren, is gebruik gemaakt van het programma OPLAT. Binnen dit programma wordt de coëfficiënt MAcc ('Accuracy of Measurement') berekend. Deze coëfficiënt vertoont qua interpretatie grote overeenkomst met de betrouwbaarheidscoëfficiënt uit de klassieke testtheorie (KTT). Deze coëfficiënt wordt in de psychometrische literatuur beschreven en als correct aangemaakt. Naast de op basis van IRT berekende MAcc is de GLB ('Greatest Lower Bound') berekend, een betrouwbaarheidsschatter op basis van KTT. In de regel komt GLB hoger uit dan de bekende (Cohen's) coëfficiënt alfa (eveneens op basis van KTT) die geldt als een lichte onderschatter van de betrouwbaarheid (hoe meer de toets afwijkt van unidimimensionaliteit, des te meer is coëfficiënt alfa een onderschatter). MAcc is doorgaans (ongeveer) gelijk aan coëfficiënt alfa en kent dus ook een lichte vorm van onderschatting die kenmerkend is voor alfa. Kortom, GLB is waarschijnlijk de coëfficiënt die de werkelijke betrouwbaarheid van de toets het beste benadert.

Conclusie:
Aan aspect B1.1 wordt het oordeel '**voldoende**' toegekend.

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

Bevindingen:
Er wordt verwezen naar de COTAN criteria voor toetsen voor minder belangrijke beslissingen (zoals de toetsen Taalverzorging). De interne consistentie betrouwbaarheid is, volgens deze criteria, voldoende bij een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80, en bij een betrouwbaarheidscoëfficiënt hoger dan 0,80 goed. Voor de toets 'Taalverzorging groep 6-8' wordt deze coëfficiënt berekend als MAcc (zie B1.1) voor de afnamemomenten M6/E6, M7/E7 en M8. Aanvullend hierop wordt de standaardmeetfout vermeld. De afnamecontext van de toets leent zich, dankzij een OPLM kalibratie, voor een gesimuleerd test-hertest onderzoek onder ideale condities. De MAcc, GLB en test-hertest coëfficiënten zijn voor de populaties M6/E6, M7/E7 en M8 vrijwel identiek voor alle onderdelen (Interpunctie, Spelling niet-werkwoorden, Grammatica en Spelling werkwoorden). Op grond van de eerdergenoemde COTAN-criteria is de betrouwbaarheid op alle normeringsmomenten (M6/E6, M7/E7 en M8) voor het deelgebied interpunctie goed te noemen (GLB variërend van 0,81 tot 0,86), voor spelling niet-werkwoorden is de betrouwbaarheid voldoende (GLB variërend van 0,76 tot 0,79), voor grammatica goed (GLB variërend van 0,81 tot 0,86) en voor spelling werkwoorden is alleen het normeringsmoment M8 voldoende (GLB= 0,70). Voor M7 en E7 is de betrouwbaarheid niet voldoende gebleken voor het onderdeel Spelling werkwoorden (de betrouwbaarheden voor de normeringsmomenten M6 en E6 worden niet vermeld voor dit onderdeel, omdat dit onderdeel daar nog niet aan bod is gekomen). Verder is de gemiddelde betrouwbaarheid van alle deelgebieden samen van de toetsen taalverzorging met een gemiddelde van 0,90 goed te noemen volgens de COTAN-criteria.

In verband met de lage betrouwbaarheid van de toetsen werkwoordenspelling M7/E7 worden leerkrachten in groep 7 aangeraden bij de interpretatie van de behaalde vaardigheidsscores de nodige voorzichtigheid te betrachten. Als er belangrijke beslissingen verbonden moeten worden aan de scores voor de deeltoets werkwoordenspelling M7/E7, wordt leerkrachten geadviseerd om gebruik te maken van de toetsen Cito Spelling werkwoorden voor groep 7 (2^{e} generatie) en de in 2017 te verschijnen toetsen Cito Spelling werkwoorden 3.0 voor groep 7.

De lage betrouwbaarheden voor M7 en E7 voor het onderdeel Werkwoordspelling bleek ook al uit de item-totaalcorrelaties. Deze zijn qua range en gemiddelde een stuk lager, in combinatie met een vrij lage gemiddelde P-waarde (M7, gem. Rit: 0,34 en E7, gem. Rit: 0,36). De verklaring hiervoor is drieledig en ligt deels in het onderwijsaanbod (werkwoordspelling is nog niet veel aan bod gekomen in groep 7), deels in de referentieniveaus (om recht te doen aan de inhoudelijke beschrijving van de morfologische spelling in het referentiekader, zijn in de taken items opgenomen die voor veel leerlingen nog te moeilijk waren) en deels in de lengte van de taken (om de leerlingen niet te veel te belasten is er consequent besloten om in alle jaargroepen slechts 20 items per deeltaak op te nemen).

Aanvullend hierop is de lokale meetnauwkeurigheid (misclassificaties) weergegeven in betrouwbaarheidstabellen. Uitgaande van de betrouwbaarheidstabellen worden twee indices voor de nauwkeurigheid van de classificaties gerapporteerd: de Accuracy plus/minus 1 niveau-index van Pilliner (1969) en de Marginal Classification Accuracy index, beide als totale samenvattende maat, en daarnaast per scoregroep eveneens de Accuracy plus/minus 1 niveau. Er is bewust voor deze maten gekozen omdat deze betrekkelijk intuïtief te interpreteren zijn, hetgeen minder geldt voor andere indices. Pilliner (1969) heeft als één van de weinigen een ambitieniveau geformuleerd voor de door hem ontwikkelde Accuracy plus/minus 1 niveau-index, nl. dat 95% van de leerlingen in een niveaugroep in werkelijkheid ook in die niveaugroep moet scoren, **of** één niveaugroep daarboven ***of*** één niveaugroep daaronder. Dit ambitieniveau is gebaseerd op de veronderstelling dat geen enkele toets perfect meet en dat er dus altijd sprake is van foutieve classificaties (alleen voor toetsen met een perfecte betrouwbaarheid van 1, is er geen sprake van misclassificaties). In dat licht bezien is de maximale accuraatheid die op het individuele niveau bereikt kan worden plus of minus één niveaugroep.

Uit de hoogte van de indices blijkt dat de laagst en de hoogst scorende leerlingen accuraat te classificeren zijn, maar dat tussen leerlingen in de niveaugroepen B, C en D, respectievelijk II, III en IV, minder duidelijk onderscheid te maken is. In het midden is de accuraatheid van de classificatie dus minder. Dit past bij één van de doelen van deze toets: signaleren welke leerlingen extra aandacht of extra uitdaging nodig hebben. Het percentage misclassificaties is bij de middelste scoregroepen het hoogst, te weten bij de scoregroep III, respectievelijk scoregroep C. Gemiddeld gezien scoort, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 82 tot 97 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* index loopt uiteen van 44 tot 62 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij beide afnamemomenten gemiddeld gezien in ruim 50 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore.

Conclusie:
De betrouwbaarheid van de toetsen Taalverzorging groep 6 tot en met 8 is 'voldoende' als aangenomen mag worden dat de toets geen zware consequenties voor de leerlingen heeft.

Op aspect B.1.2. wordt aan de toets 'Taalverzorging groep 6 tot en met 8' 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:
Zoals uit de algemene beschrijving hierboven blijkt, is de Handleiding voor de leerkracht zeer leesbaar en volledig. De Handleiding gaat uitgebreid in op het voorbereiden, afnemen, nakijken en verwerken van de toetsen en daarmee op hun plek in het onderwijsleerproces. De keuze voor en gebruiksmogelijkheden van de toetsversies voor diverse deeldoelgroepen en meetmomenten in groep 6-8 wordt grondig toegelicht. Daarnaast is aandacht voor de mogelijkheid tot categorieënanalyse, interpreteren en analyseren van

resultaten op leerling- en groepsniveau, en voor communicatie met leerling en ouders over de resultaten. Voor de interpretatie is het belangrijk op te merken dat hier een opmerking opgenomen kan worden over de betrouwbaarheid van de toetsen werkwoordspelling M7/E7 en dat leerkrachten worden aangeraden tevens te kijken naar de toetsen Cito Spelling werkwoorden voor groep 7. Deze opmerking is opgenomen in de samenvatting van de wetenschappelijke verantwoording het is echter niet aannemelijk dat leerkrachten deze lezen. Deze opmerking opnemen in de handleiding zou daarom verstandig zijn. Cito B.V. heeft inmiddels toegezegd deze opmerking op te nemen in de tweede druk van de Handleiding.

De toetsen Taalverzorging voor groep 6, 7 en 8 bestaan per jaargroep telkens uit drie (M6/E6) of vier taken (M7/E7, M8); elke taak bevat in totaal 20 meerkeuzeopgaven. Afname is in principe klassikaal. De wijze waarop een leerkracht de toetsen 'op maat' in kan zetten en evalueren wordt uitgebreid toegelicht en tevens verbonden met de rapportage van de reguliere afname en eventueel naar een handelingsplan en schoolplan.

Bij de toetsconstructie is aantoonbaar aangesloten bij diverse relevante bronnen, waaronder de tussendoelen en leerlijnen en de Referentieniveaus Taalverzorging ('Referentiekader taal en rekenen'). In de wetenschappelijke verantwoording is de keuze en verdeling in subdomeinen, categorieën en soorten opgaven onderbouwd.

De toetsontwikkelaars leveren met de toets een uitwerking voor groep 6, 7 en 8 van het PO en SO, van de volgende onderdelen van schriftelijke taalvaardigheid:

- Spelling van werkwoorden
- Spelling van niet-werkwoorden
- Interpunctie
- Grammatica

De enige uitzondering vormt toetsing van werkwoordspelling, die in groep 6 te weinig structureel aan bod komt om dan al opgenomen te worden in de LVS-toetsing. De toetsen kunnen in de groepen 6 en 7 halverwege of aan het einde van het schooljaar worden afgenomen. Voor groep 8 is alleen toetsing voor gebruik halverwege het schooljaar ontwikkeld. Voor alle afnamemomenten is een normering beschikbaar.

De toelichting op de inhouden van elk van de toetsonderdelen is uitgebreid en helder. Verhelderend voor de gebruiker is tevens dat vervolgens de diverse opgavenvormen worden gepresenteerd en toegelicht.

De vele verschillende opgaventypen en daaraan gekoppelde vraagstellingen die kunnen voorkomen, zorgen er wel voor dat de leerlingen gedegen moeten worden voorbereid op het belang van het goed lezen van de opgave. Door de aard van het te toetsen domein is de vraagstelling wel gekunsteld: de zinnen zijn doorgaans niet representatief voor zinnen die leerlingen in teksten tegenkomen. De moeilijkheidsgraad is, gezien het niveau van de doelgroep, aanvaardbaar. Verder laat het antwoordmodel geen ruimte voor interpretatie. Het niveau van de teksten en vraagstellingen in het toetsmateriaal geven geen aanleiding tot fundamentele vragen over of twijfel aan de validiteit.

Merk op dat de beoordeling van dit aspect zich hieronder beperkt tot het statistisch/psychometrisch onderzoek dat verricht is inzake validiteit en dan specifiek begripsvaliditeit.

De toets Taalverzorging groep 6-8 is niet bedoeld voor voorspellend gebruik. Daarmee is de criteriumvaliditeit niet van toepassing. De (psychometrische) begripsvaliditeit wordt uitgewerkt in unidimensionaliteit, itemkwaliteit, itembias, convergente en discriminante (of divergente) validiteit en in verschillen tussen relevante subgroepen.

De resultaten van de uitgevoerde kalibratie maken het aannemelijk dat er sprake is van unidimensionaliteit. Dit betekent dat met elke willekeurige subset van items uit de gekalibreerde itembank dezelfde onderliggende (latente) vaardigheid 'Taalverzorging groep 6-8' per deeltoets kan worden vastgesteld (de vier deelgebieden worden beschouwd als vier unidimensionale deelvaardigheden die tezamen de vaardigheid 'Taalverzorging' vormen). Dit wordt tevens bevestigd door de nauwkeurigheid van de itemparameterschattingen (de constante 'c' uit het COTAN-systeem is voor alle items <= 0.20, hetgeen als 'goed' te kwalificeren is en daarmee duidt op een hoge nauwkeurigheid van de itemparameterschattingen en dientengevolge op een hoge itemkwaliteit).

Een andere indicator voor de itemkwaliteit is de gemiddelde moeilijkheidsgraad (p-waarde uit de KTT) van de toetsen. Deze ligt op het (vooraf) gewenste niveau, namelijk voor interpunctie tussen 0,67 (M6) en 0,75 (E6), voor spelling niet-werkwoorden tussen 0,68 (M6) en 0,75 (E6), voor grammatica tussen 0,67 (M6) en 0,71 (E7) en voor spelling werkwoorden tussen 0,59 (M8) en 0,65 (E7). De gemiddelde moeilijkheidsgraad voldoet daarmee aan het gestelde doel, namelijk een optimaal onderscheidend vermogen bij de groep met een lage of gemiddelde vaardigheid, terwijl de toetsen niet als bijzonder moeilijk zullen worden ervaren door de doorsnee leerling. De moeilijkheidsgraad van de afzonderlijke opgaven kent een goede spreiding; er zijn zowel moeilijke als gemakkelijke opgaven in de toetsen opgenomen om zodoende rekening te houden met respectievelijk het plafondeffect en het vloereffect. Ook zijn gemakkelijke opgaven opgenomen vanwege motivationele factoren (succeservaring opdoen). Het merendeel (90%) van de opgaven is niet te moeilijk (p-waarde < 0,20) of te makkelijk (p-waarde > 0,80).

De gemiddelde samenhang tussen item- en totaalscore (Rit-waarden uit de KTT) als nog een andere indicator voor de (gemiddelde) itemkwaliteit voldoet eveneens aan normwaarden van het COTAN-beoordelingssysteem. Geen enkele opgave heeft een lagere Rit-waarde dan 0,25. De laagste waarden treffen we aan bij de drie opgaven van de toetsen Spelling werkwoorden, maar de waarden vallen wel binnen de range voldoende. De gemiddelde Rit-waarden zijn voor de toetsen te kenschetsen als 'goed' (gemiddelde Rit-waarden > 0,30).

In het onderzoek naar itembias (differentieel functioneren) is er bij slechts één item (in de toets M7/E7) sprake van DIF (Differential Item Functioning) met betrekking tot sekse. De gemiddelde scores op Taalverzorging naar sekse laten zien meisjes als groep iets hoger scoren dan jongens (op het 1% significantieniveau). Deze bevinding sluit aan bij het gegeven dat meisjes bij talige toetsonderdelen over het algemeen enigszins in het voordeel zijn. De verschillen zijn echter klein.

De constructvaliditeit is uitgewerkt in convergente en divergente validiteit door de samenhang (i.e., intercorrelaties) te onderzoeken tussen de toetsen Taalverzorging en andere taaltoetsen uit het Cito Volgsysteem primair onderwijs van de tweede generatie. De LVS-toetsen van de tweede generatie hebben betrekking op alle afnamemomenten tussen M6 en M8. Al deze toetsen uit de tweede generatie van het LVS zijn door de Cotan

op alle relevante onderdelen (criteriumvaliditeit is niet van toepassing) met een goed of voldoende beoordeeld.

Uit deze gegevens blijkt dat de scores op de toetsen Taalverzorging sterk samenhangen (i.e., aanwijzing voor convergente validiteit) met scores op meer technische taalvaardigheidsonderdelen, zoals spelling en technisch lezen (leestempo), en minder (i.e., aanwijzing voor divergente validiteit) met scores op andere, meer semantische onderdelen van taalvaardigheid, zoals begrijpend lezen. Ook de correlatie met rekenen-wiskunde bleek naar verwachting matig te zijn, omdat het een geheel andere vaardigheid betreft dan de deelvaardigheden van Taalverzorging. Samenvattend kan dus gesteld worden dat de samenhangen van de toetsen Taalverzorging met andere toetsscores conform de verwachtingen zijn. De data geven aan dat er gemeten wordt wat men beoogt te meten, namelijk deelvaardigheden van taalverzorging en vormen dus een ondersteuning voor de begripsvaliditeit van de toets Taalverzorging.

Ook is gekeken naar de samenhang tussen de verschillende deelvaardigheden van Taalverzorging. De verwachting was dat de samenhangen tussen de deelvaardigheden matig zijn. Aanvankelijk (in groep 6) bleek de samenhang inderdaad matig te zijn, hetgeen een ondersteuning is voor de beslissing om taalverzorging niet op te vatten als een unidimensionaal concept maar om een indeling in vier deelgebieden aan te houden, met ieder hun eigen itembank. Bij een domein als bijvoorbeeld Rekenen is de samenhang tussen de deelvaardigheden veel groter en treffen we in de bovenbouw doorgaans correlaties aan van ruim boven 0,90, hetgeen rechtvaardigt om Rekenen in tegenstelling tot Taalverzorging op te vatten als één-unidimensionale vaardigheid. Op basis van de correlaties kon echter ook worden aangetoond dat de samenhang tussen de vier deelgebieden toeneemt doordat de deelvaardigheden in het leerproces steeds beter geïntegreerd raken naarmate dat proces vordert. De leerlingen gaan de verschillende deelvaardigheden dus steeds meer integraal toepassen naarmate het leerproces vordert. Taalverzorging neigt zich dus langzamerhand te ontwikkelen tot een meer unidimensionale vaardigheid (de deelvaardigheden van taalverzorging worden alsnog geïsoleerd aangeboden in het onderwijs), al lijkt het laatste meetmoment voor deze reeks toetsen (M8) nog net te vroeg te komen om van een geïntegreerde unidimensionale vaardigheid te kunnen spreken als basis voor de toets(en).

Als laatste werden verschillen tussen relevante subgroepen (naar leeftijd, sekse en leerklinggewicht) gepresenteerd. De resultaten bleken aan te sluiten bij de verwachtingen die op grond van theoretische inzichten en eerder onderzoek konden worden geformuleerd (o.a. dat meisjes doorgaans bij 'talige vaardigheden' hoger scoren dan jongens, specifiek op spellingstoetsen) en vormen daarmee extra ondersteuning voor de begripsvaliditeit van de toetsen. Wat betreft de verschillen tussen relevante subgroepen scoren jongere leerlingen naar verwachting iets beter dan oudere leerlingen en scoren meisjes iets hoger dan jongens.

Conclusie:
Aan aspect V1. wordt het oordeel '**voldoende**' toegekend.

Het volg-aspect

<u>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?</u>

Bevindingen:
Het algemene (inhoudelijke) uitgangspunt voor de toets Taalverzorging groep 6-8 is dat ieder van de (latente) vaardigheden van de deelvaardigheden Taalverzorging groep 6-8 kan worden opgevat als een unidimensionaal continuüm en dat elke leerling kan worden voorgesteld als een punt op dit continuüm. Uit het kalibratieonderzoek blijkt dat de items van de deelvaardigheden passen bij het gehanteerde IRT model en dat het model ook past voor de toetsen M6/E6, M7/E7 en M8 als geheel. Dit betekent dat er sprake is van één unidimensionale vaardigheidsschaal waar items en leerlingen op afgebeeld kunnen worden.

Afhankelijk van het aantal items dat een leerling goed maakt, wordt er een vaardigheidsscore toegekend. Jongere leerlingen scoren iets beter dan oudere leerlingen. Tevens scoren meisjes iets hoger dan jongens.

Voor Taalverzorging groep 6-8 is een nieuwe vaardigheidsschaal ontwikkeld, waarop alle uitgebrachte en nog uit te brengen toetsen uit het Cito Volgsysteem primair en speciaal basisonderwijs Taalverzorging groep 6-8 worden gekalibreerd.

Conclusie:
Aan aspect VA1.1. wordt 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 hoofdstuk 7 van de handleiding ('Communiceren over toetsresultaten met leerling en ouders') wordt beschreven hoe er met de verschillende gebruikers over de toetsresultaten kan worden gecommuniceerd. Hierin wordt onderscheid gemaakt tussen 'niveau' en 'groei', wat wordt onderbouwd met diverse rapportage mogelijkheden.

In de wetenschappelijke verantwoording wordt toegelicht hoe de toetsen ingezet kunnen worden om de ontwikkeling van leerlingen te volgen in de tijd, namelijk door het toetsresultaat van een leerling te vergelijken met andere leerlingen en door het toetsresultaat van een leerling te vergelijken met diens andere toetsresultaten. Voor alle vergelijkingen geldt dat uitspraken over de voortgang van leerlingen gerelativeerd moeten worden vanwege de (on)betrouwbaarheid van de toetsen. Door betrokkenen bij de toetsen Taalverzorging groep 6-8 moet beseft worden dat vaardigheidsgroei zich langzaam in de tijd voltrekt.

Conclusie:
Aan aspect VA1.2. wordt 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?

Bevindingen:
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). Voor ouders is vooral het leerlingrapport of alternatief leerlingrapport informatief omdat deze rapporten van hun kind individueel de vaardigheid en de groei weergeven.

In de Handleiding wordt in hoofdstuk 7 aandacht besteed aan de wijze waarop met ouders over de toetsresultaten gecommuniceerd kan/moet worden. Vooral wordt daarbij gewezen op het leerlingrapport waarin zowel het niveau van de leerling als de progressie van de leerling numeriek en grafisch gepresenteerd worden.

Daarnaast wordt de leraar 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 (brengen één of twee keer per jaar in beeld wat de leerling weet en kan onderscheid maken tussen leerlingen die de leerstof maar net beheersen en leerlingen die wat extra's aankunnen) en methode-gebonden (een groot deel van de leerlingen zou deze (bijna) foutloos moeten kunnen maken) toetsen duidelijk maken en erop wijzen dat deze toetsen leerlingen anders (kunnen) beoordelen. De informatie biedt goede handvatten voor de gesprekken met ouders. In hoofdstuk 8 worden veelgestelde vragen behandeld die weliswaar voor de leraren bestemd zijn maar waar de antwoorden voor een deel ook informatief zijn tijdens bijvoorbeeld de tienminutengesprekken.

De categorieënanalyse voor de vier deelvaardigheden interpunctie, spelling niet-werkwoorden, grammatica en spelling werkwoorden die worden weergegeven in bijlage 4, bedoeld als een eerste hulpmiddel waarmee de toetsresultaten in een breder perspectief kunnen worden geplaatst, kunnen ook behulpzaam zijn in de communicatie naar ouders toe. De categorieënanalyse is bedoeld als hulpmiddel voor de leerkracht om na te gaan of de leerling, gegeven zijn vaardigheidsniveau, evenwichtig presteert op elk van de verschillende deelgebieden van de toetsen Taalverzorging. In de grafische presentatie bij de categorieënanalyse kan een leerkracht zien of een leerling bij een bepaald deelgebied beter of zwakker scoort dan verwacht. Naast een 'categorieënanalyse leerling' is er ook een zogenoemde 'categorieënanalyse groep'. In deze laatste rapportage staan eerst per leerling alle resultaten uit de 'categorieënanalyse' overzichtelijk onder elkaar voor de hele groep. Vervolgens staat in een tabel aangegeven hoeveel leerlingen uit die groep per categorie beneden en hoeveel leerlingen boven verwachting scoren, inclusief de gemiddelde afwijking naar beneden/boven.

Over de interpretatie van toetsresultaten is ook een folder ouderinformatie beschikbaar die men via de website van het Cito kan downloaden.

Conclusie:
Op aspect I1.1 wordt aan de toetsen Taalverzorging groep 6-8 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	V1.1	**Voldoende**
Volg-aspect	VA1.1	**Voldoende**
	VA1.2	**Voldoende**
Inzicht in leervorderingen	I1.1	**Voldoende**

4. Literatuurlijst

- Engelen, R., Roumans, P., Keizer, M. & Jongen, I. (2016). *Wetenschappelijke verantwoording van Taalverzorging groep 6 tot en met 8.* Arnhem: Cito B.V.
- Cito (2015). *Cito Volgsysteem primair en speciaal onderwijs. Taalverzorging groep 6 tot en met 8*. Arnhem: Cito