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 toetsen Begrijpend lezen 3.0 maken deel uit van de derde generatie toetsen van het Cito Volgsysteem primair en speciaal onderwijs en zijn bedoeld voor leerlingen in groep 8 van het basisonderwijs. De toetsen zijn ook geschikt voor leerlingen in het speciaal basisonderwijs en het speciaal onderwijs cluster 2 en 4. Het betreft papieren toetsen.
Onderstaande beschrijving is gebaseerd op de wetenschappelijke verantwoording.

Meetpretentie

De toetsen Begrijpend lezen 3.0 beogen de vaardigheid in het begrijpen van geschreven teksten te meten. De toetsen zijn bedoeld om vast te stellen hoe goed een leerling geschreven teksten kan begrijpen en hoe de vaardigheid in begrijpend lezen van de leerling zich in de loop van de jaren ontwikkelt.

Doelgroep

De toetsen Begrijpend lezen 3.0 groep 8 zijn bedoeld voor leerlingen in groep 8 van het basisonderwijs. De toetsen zijn ook geschikt voor leerlingen in het speciaal basisonderwijs en het speciaal onderwijs cluster 2 en 4. Voor deze groepen speciale leerlingen zijn geen afzonderlijke normen vastgesteld. De toetsresultaten van deze leerlingen worden geïnterpreteerd met behulp van de gemiddeld vaardigheidsscores voor leerlingen uit het reguliere onderwijs.

Gebruiksdoel en functie

De toetsen Begrijpend lezen 3.0 hebben twee doelen:

- Het leesvaardigheidsniveau van zowel individuele leerlingen als groepen leerlingen te beoordelen via een vergelijking van de behaalde scores met de scores van een referentiegroep oftewel niveaubepaling.
- De ontwikkeling van de leesvaardigheid van zowel individuele leerlingen als groepen leerlingen door de leerjaren heen te volgen oftewel progressiebepaling.

Inhoudelijke theoretische inkadering:

De inhoud van de toetsen Begrijpend lezen 3.0 is gebaseerd op het domein Lezen, waarbij onderscheid is gemaakt tussen het lezen van zakelijke teksten en het lezen van fictionele, narratieve en literaire teksten, 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 lezen 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 zowel de theoretische basis als voor de indeling van de vaardigheden. De indeling van de vaardigheden is mede gebaseerd op basis van analyse van methoden voor begrijpend lezen die veel gebruikt worden in het basisonderwijs.

Inhoud van het toetspakket

Het toetspakket Begrijpend lezen 3.0 groep 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),
  - algemene aandachtspunten voor het schoolplan (hfdst 5),
  - inhoudsverantwoording (hfdst 6),
  - communiceren over toetsresultaten met leerling en ouders (hfdst 7),
  - achtergrondinformatie en veel gestelde vragen (hfdst 8) en
  - enkele bijlagen
- Toets:
  - Tekstboekje B8/M8
  - Opgavenboekje B8/M8
- Afnamekaart
- Nakijkkaart
- Tabel voor het bepalen van de vaardigheidsscore en –niveau
- Wetenschappelijke verantwoording

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor (reeksen van) toetsen uit leerlingvolgsystemen (LVS)', 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 Jennifer Roubiës MSc (secretaris).

De kwaliteit van de dataverzameling

S1 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 (ongeveer 50 procent) te baseren op gegevens uit het zogenaamde embedded field normeringsonderzoek en (voor ongeveer 50 procent) op gegevens uit Cito dataretour die afkomstig zijn van de tweede generatie toetsen LVS-toetsen. Bij de normering van de derde generatie LVS-toetsen is rekening gehouden met de variabelen regio, urbanisatiegraad, schooltype (stratum) en sekse, waarbij de eerste drie variabelen op schoolniveau zijn gedefinieerd en de laatste variabele op leerlingniveau. Bij regio is uitgegaan van de vier landsdelen/regio's van de CBS-indeling (noord, oost, west, zuid). Bij urbanisatiegraad is ervoor gekozen om de indeling naar vijf niveaus, die gebruikelijk is bij het CBS, te reduceren tot een tweedeling in enerzijds niet tot matig verstedelijkt (platteland) en anderzijds sterk tot zeer sterk verstedelijkt (stad). Een dergelijke tweedeling blijkt in de praktijk goed te volstaan (cf. Van Boxtel en Hemker, 2009). Bij schooltype is uitgegaan van de formatiegewichten volgens OCW. Hierin worden drie niveaus onderscheiden die gebaseerd zijn op het opleidingsniveau van de ouders. Bij sekse is een tweedeling gemaakt naar jongens en meisjes.

Op pagina 48-50 van de Wetenschappelijke Verantwoording (hierna: WV) 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 van gegevens uit DUO te vergelijken met de steekproefverdelingen van afnamemomenten B8 en M8. Tabel 4.5 en 4.6 presenteren de resultaten van de representativiteitsanalyses van respectievelijk B8 en M8. Voor de achtergrondvariabele schooltype zaten er in de normeringssteekproef voor afnamemoment B8 geen scholen met 25% of meer achterstandsleerlingen. Daarom is besloten om voor afnamemomemt B8 een andere indeling voor schooltype te gebruiken (schooltype 2): scholen met minder én scholen met meer dan 15% achterstandsleerlingen. Tabel 4.6 laat zien dat voor afnamemoment M8 de steekproefverdeling weinig afwijkt van de populatieverdeling. Tabel 4.5 laat zien dat er voor afnamemoment B8 afwijkingen zijn van landelijke verdelingen wat betreft regio en stratum.

De afwijkingen zijn ook statistisch getoetst. De gegevens van deze statistische toetsen worden getoond in de tabellen 4.7 en 4.8 voor respectievelijk B8 en M8. Tabel 4.7 laat zien dat normeringssteekproef B8 een redelijk goede afspiegeling

vormt van de populatie en alleen het effect (wortel uit quotiënt van chi-kwadraat en steekproeflengte) voor regio gemiddeld is (ϕ = 0.552). Statistische weging van de resultaten is daarom niet nodig. Tabel 4.8 laat zien dat voor afnamemoment M8 de chi kwadraat-waarden voor regio en sekse significant zijn. We zien dat de effectgrootte voor de variabele regio rond de 0.50 ligt (ϕ = 0.559), wat een gemiddeld effect weerspiegelt (cf. Cohen, 1988). Voor sekse is de effectgrootte klein (ϕ = 0.164). Ook voor de normeringssteekproef M8 geldt dat deze een redelijk goede afspiegeling vormt van de populatie en dat daarom statistische weging van de resultaten wederom niet nodig is.

In tabel 4.4 van de WV worden de aantallen leerlingen weergegeven die meegenomen zijn bij de normering. Voor M8 betreft dat 1.754 leerlingen (788 leerlingen uit de steekproef en 966 leerlingen via dataretour) afkomstig van 121 scholen en voor B8 963 leerlingen uit de steekproef afkomstig van 29 scholen. Voor B8 was aanvulling met gegevens uit dataretour niet mogelijk.

Uit de ruwe scores van de individuele 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 dus niet alleen informatie over de geschatte vaardigheid, maar ook over de onzekerheid die bij die schatting hoort (Keuning et al., 2015). De normering werd vervolgens gebaseerd op de 'plausible values' van de leerlingen in de normeringssteekproef. De 'plausible values' voor de afnamemomenten B8 en M8 bleken een normale verdeling te vormen. Op basis van deze scoreverdeling werden de percentielen berekend die horen bij de vaardigheidsverdelingen A t/m E en I t/m V zoals beschreven in hoofdstuk 2 van de WV. De schoolverdeling werd bepaald met het intercept-only multilevel model met een gemiddelde per school en een variantie op school- en leerlingniveau. Dit model werd geschat via een bootstrap procedure. Dit betekent dat het multilevel model meerdere keren wordt geschat, steeds op basis van een andere selectie van scholen en leerlingen uit de normeringssteekproef. Ondanks dat de percentielen van de normgegevens op schoolniveau dichter bij elkaar kwamen te liggen dan in de leerlingverdeling, waren de afstanden groot genoeg om scholen zinvol te classificeren in de verschillende niveaus.

Conclusie:
De steekproeven zijn representatief, zijn adequaat gestratificeerd naar regio, urbanisatiegraad, schooltype en sekse 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 toetsen worden afgenomen. Op aspect S1 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het volgende oordeel toegekend: '**voldoende**'.

S2 In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?

Bevindingen:
Bij de ontwikkeling van de toets Begrijpend lezen 3.0 groep 8 is geprobeerd om door middel van embedded field designs bias in de normen te vermijden door de afnamesituatie waarin de toets wordt afgenomen tijdens het normeringsonderzoek zoveel mogelijk te laten lijken op de afnamesituatie na uitgave van de toets. In een embedded field design lopen nieuw ontwikkelde items voor de derde generatie mee in de al bestaande en op scholen toegepaste toetscyclus. Bij de leerlingen was onbekend welke taken de nieuwe opgaven bevatten. Tevens was voor de leerlingen onbekend dat de gegevens ook voor onderzoeksdoeleinden werden gebruikt. Voor deze opzet werd gekozen opdat motivatie-effecten de verzamelde gegevens voor het normeringsonderzoek zo min mogelijk zouden beïnvloeden. Een belangrijk tweede voordeel van deze aanpak is dat de normeringssteekproef M8 aangevuld kan worden met resultaten uit dataretour van de tweede generatie LVS-toetsen (zie Keuning et al., 2015).

Het normeringsonderzoek gaat uit van een onvolledig maar 'verbonden' dataverzamelingsdesign, waartoe een longitudinale opzet werd gebruikt. In totaal lieten scholen hun leerlingen op afnamemoment M8 vier taken maken; een taak uit LVS 2.0 en drie taken met nieuw materiaal voor de derde generatie. In figuur 4.3 wordt het embedded field design weergegeven voor toetsafname M8, waarvan vijf toetsversies zijn afgenomen. Op afnamemoment M8 maakten de leerlingen volgens het design een taak uit de M8-toets LVS 2.0. Daarnaast maakte elke leerling drie taken met nieuwe opgaven voor LVS 3.0 (M8 deel 1, M8 deel 2 en M8 deel 3). Sommige leerlingen maakten twee taken met nieuwe opgaven voor LVS 3.0 en daarnaast een ankertaak 'referentieset' om de prestatiestandaard van de referentie-niveaus over te kunnen brengen. De taak E7 anker in het design bevat opgaven uit de beoogde selectie van E7. Er zijn ook twee 'reservetaken' meegenomen in het normeringsonderzoek. Deze opgaven zouden in de uiteindelijke uitgave alleen worden gebruikt indien er onverwachte problemen naar voren kwamen met betrekking tot de beoogde taken voor de uitgave M8.

Het design in figuur 4.3 laat zien dat de (toets)boekjes ('booklets') stevig verankerd zijn. Het deel M8 LVS 2.0 vormt een stevig anker tussen de toetsboekjes voor afnamemoment M8. Binnen afnamemoment M8 werd er geankerd door middel van de taken met nieuw ontwikkeld materiaal voor LVS 3.0. Tevens werd er geankerd tussen de beoogde toetsen voor LVS 3.0. Er is dus ook rekening gehouden over de toetsmomenten heen.

In het normeringsonderzoek M8 zijn 202 items voorgelegd aan 875 leerlingen medio groep 8, verdeeld over 5 boekjes volgens het design in figuur 4.3. Elk boekje bestond uit 100 tot 102 opgaven verdeeld over 4 taken. De 25 opgaven in de LVS 2.0 taak werden door alle 875 leerlingen gemaakt. De overige opgaven kwamen in twee of drie boekjes voor en werden gemiddeld door 390 leerlingen gemaakt. Bij alle opgaven werd aan de eis dat deze minimaal bij 150 leerlingen moesten zijn afgenomen, voldaan.

Op grond van signalen uit het veld is in latere instantie besloten om ook een B8-normering vast te stellen. Scholen bleken er namelijk behoefte aan te hebben om al in oktober/november in groep 8 te toetsen, om zodoende de B8-normering te kunnen gebruiken voor uitstroomadvies. Cito heeft daarom besloten om alsnog (in oktober 2018) een B8-normeringsonderzoek uit te voeren. Op deze manier is er gezorgd voor één toets Begrijpend lezen voor groep 8 (i.e., toets B8/M8) die naar wens in oktober/november (afnamemoment B8) of januari/februari (afnamemoment M8) af te nemen is.

In het normeringsonderzoek B8 maakten leerlingen van groep 8 de inmiddels al uitgegeven toets voor groep 8. Het design van het onderzoek voor B8 bestond dus uit 3 taken (in totaal 73 items) van de definitieve toets voor groep 8. Deze toets werd in oktober 2018 afgenomen door scholen en de data werd aan Cito geleverd in de vorm van dataretour. Om deelnemers te kunnen werven voor het normeringsonderzoek B8 zijn scholen aangeschreven. 15 scholen hadden zich opgegeven, met 515 leerlingen. Met behulp van dataretour konden de data worden aangevuld tot 920 leerlingen, omdat al meer scholen de uitgebrachte toets voor groep 8 in oktober al hadden afgenomen. Voor de normering B8 werden ook de leerlingen meegenomen die een makkelijkere toets (bijv. E7) gemaakt hadden; dit zijn veelal de iets minder vaardige leerlingen. In totaal bestond de normeringssteekproef voor B8 dan uit 963 leerlingen verdeeld over 29 scholen (zie ook tabel 4.4). De opgaven in het normeringsonderzoek B8 waren al gekalibreerd, omdat de leerlingen immers de al uitgegeven toets voor groep 8 hadden gemaakt. Alles wat hierboven is aangegeven over de kalibratie van afnamemoment M8 (zie figuur 4.3) is daarom ook van toepassing op afnamemoment B8.

Uit het kalibratieonderzoek blijkt dat voor de toets B8/M8 zowel de items, gegeven de M- en S-toetsen, als de gehele toetsen, gegeven de R1c-toetsen (R1c = 951.927, hetgeen minder is dan anderhalf maal het aantal vrijheidsgraden van 822), passen bij het itemresponstheoriemodel OPLM (met dit statistische model zijn de moeilijkheidsparameters en discriminatie-indices van de items geschat). Ook de zogenoemde constante 'c' (deze constante geeft de relatie weer tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie) uit het COTAN-Beoordelingssysteem bevestigt deze conclusie (range = 0.06 – 0.29 en gemiddelde = 0.14, d.w.z. kleiner dan 0.30 en daarmee ruim voldoende volgens de criteria die de COTAN hanteert). Dit betekent dat er sprake is van een eendimensionale vaardigheidsschaal waar items en leerlingen op afgebeeld kunnen worden. Op grond van de psychometrische gegevens uit de kalibratie (en inhoudelijke criteria) zijn de definitieve selecties van items gemaakt voor de uitgave van de toets Begrijpend lezen 3.0 groep 8 (afnamemomenten B8 en M8).

Conclusie:
Het onvolledige maar 'verbonden' dataverzamelingsdesign is adequaat. Op aspect S2 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel **'voldoende'** toegekend.

Normering

N1.1 Is de standaardbepalingsmethode gemotiveerd en op de juiste wijze uitgevoerd?

Bevindingen:
Dit criterium heeft betrekking op absoluut normeren en is dus niet van toepassing (n.v.t.) op de toets Begrijpend lezen 3.0 groep 8, omdat het hier gaat om relatief normeren.

Conclusie:
n.v.t.

N1.2 Zijn de beoordelaars/vakdeskundigen/experts naar behoren geselecteerd en getraind?

Bevindingen:
Dit criterium heeft betrekking op absoluut normeren en is dus niet van toepassing (n.v.t.) op de toets Begrijpend lezen 3.0 groep 8, omdat het hier gaat om relatief normeren.

Conclusie:
n.v.t.

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

Bevindingen:
Dit criterium heeft betrekking op absoluut normeren en is dus niet van toepassing (n.v.t.) op de toets Begrijpend lezen 3.0 groep 8, omdat het hier gaat om relatief normeren.

Conclusie:
n.v.t.

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:
In tabel 4.9 van de WV wordt voor B8 en M8 de normtabel met relatieve normen op leerlingniveau gepresenteerd. Voor aantallen leerlingen en scholen zie S1. 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 en P90. Met behulp van deze percentielen kunnen de twee niveau-indelingen (asymmetrisch: A t/m E en symmetrisch: I t/m V) opgesteld worden.

In tabel 4.12 van de WV wordt voor B8 en M8 de normtabel met relatieve normen op schoolniveau gepresenteerd. Voor aantallen leerlingen en scholen zie S1. 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 en P90.

Conclusie:
De normgroepen zijn groot genoeg (zie S1). Op aspect N2.1 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel **'voldoende'** toegekend.

N2.2 Zijn de normgroepen representatief?

Bevindingen:
De representativiteit van de steekproeven werd in S1 en S2 besproken en daar werd geconstateerd dat de steekproeven representatief waren voor regio, urbanisatiegraad, schooltype en sekse. Voor Begrijpend lezen 3.0 groep 8 geldt dat de normen geldig zijn tot en met 2027. Daarnaast monitort Cito periodiek de normering. Jaarlijks wordt aan de hand van representatieve afnamedata nagegaan of er systematisch verschuivingen in het prestatieniveau plaatsvinden. Indien nodig wordt de normering aangepast.

Conclusie:
Op aspect N2.2 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel **'voldoende'** toegekend.

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
In hoofdstuk 5 van de WV wordt beschreven hoe, gebruikmakend van het feit dat de items uit de toets geschaald zijn met behulp van de OPLM software, 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 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel **'voldoende'** toegekend.

B2 Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets genomen worden?

Bevindingen:
Voor B8 en M8 zijn drie betrouwbaarheidsgegevens berekend: standaardmeetfout, MAcc en een gesimuleerde test-hertest betrouwbaarheidscoëfficiënt. Voor afnamemoment B8 zijn de gegevens gelijk aan respectievelijk 3.58, 0.903 en 0.903, en voor afnamemoment M8 gelijk aan respectievelijk 3.55, 0.905 en 0.905 (zie

tabel 5.1). De auteurs van de WV verwijzen naar het beoordelingssysteem van de COTAN waar voor tests die geen zware consequenties voor leerlingen hebben, zoals de toetsen Begrijpend lezen 3.0 groep 8, een betrouwbaarheidscoëfficiënt van meer dan 0.80 als 'goed' aangemerkt wordt.

Naast klassieke betrouwbaarheidscoëfficiënten is ook de lokale betrouwbaarheid en de meetnauwkeurigheid onderzocht. De meetfout bleek het kleinst te zijn in de lagere en gemiddelde vaardigheidsregionen. 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 B8 en E8 (zie tabel 5.2a en tabel 5.2b). Uitgaande van de betrouwbaarheidstabellen worden twee indices voor de nauwkeurigheid van de classificaties gerapporteerd: de plus/minus 1 niveau-index en de marginal classification accuracy index. De eerste index stelt als ambitieniveau dat 95% van de leerlingen in een scoregroep in werkelijkheid ook in die scoregroep moet scoren, of één scoregroep daarboven of één scoregroep daaronder (Pilliner, 1969). In de tabellen 5.2a en 5.2b 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 misclassificaties. In dat licht bezien is de maximale accuraatheid die op het individuele niveau bereikt kan worden plus of minus één scoregroep. De tweede 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 index rond 0.75 – 0.80 uit te komen. In de praktijk liggen de waarden vaak tussen 0.60 en 0.70. Tabel 5.3 toont de samenvattende indices toets B8/M8 op afnamemomenten begin en medio groep 8.

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. Gegeven het gebruiksdoel van de toetsen kan daarom geconcludeerd worden dat de classificaties voldoende betrouwbaar zijn, maar dat er wel rekening mee moet worden gehouden dat er altijd sprake zal zijn van misclassificaties ter grootte van veelal maximaal 1 niveau.

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

Conclusie:
De betrouwbaarheid van de toetsen Begrijpend lezen 3.0 groep 8 is 'voldoende' als aangenomen mag worden dat de toets geen zware consequenties voor de leerlingen heeft en er rekening mee gehouden wordt dat er altijd sprake zal zijn van misclassificaties ter grootte van veelal maximaal 1 niveau. Op basis van het

voorgaande wordt op aspect B2 aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel '**voldoende**' toegekend.

Validiteit

V1 Inhoudsvaliditeit: 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:
De gekozen teksten in het materiaal geven een passende en voor leerlingen herkenbare range tekstsoorten en inhouden, uit verschillende genres. Ook de soorten toetsopgaven en antwoordmogelijkheden geven over de grote linie op relevante, objectieve en efficiënte wijze een beeld van de vaardigheid in begrijpend lezen van leerlingen.

Conclusie:

Op aspect V1 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel 'voldoende' toegekend.

V2 Constructvaliditeit: Meet de toets in zijn geheel datgene wat hij beoogt te meten?

Bevindingen:
De vaardigheid van leerlingen in groep 8 op het vlak van lezen met begrip wordt door de gekozen materialen en vragen, onderbouwd door de verantwoording van de constructie van de toets, inderdaad getoetst.

Vanuit psychometrisch oogpunt wordt aan bovenstaande toegevoegd:

In de WV wordt verslag gedaan van onderzoek naar verschillende aspecten van constructvaliditeit: unidimensionaliteit, itemkwaliteit, itembias, soortgenotenvaliditeit in de vorm van convergente en divergente validiteit en verschillen tussen relevante subgroepen. Voor unidimensionaliteit is verwezen naar het kalibratieonderzoek van de toetsen in hoofdstuk 4 waarvan de conclusie is dat het aannemelijk is dat er sprake is van unidimensionaliteit en dat de items één latente trek of construct meten die we, gezien ook de uitkomsten van het hierna genoemde onderzoek, kunnen aanduiden als 'vaardigheid begrijpend lezen'. Dit betekent dat met elke willekeurige subset van items dezelfde onderliggende vaardigheid kan worden vastgesteld. Hiermee is dus voldaan aan de noodzakelijke voorwaarde voor constructvaliditeit. De hoge intercorrelaties tussen de drie verschillende inhoudelijke subvaardigheden - Begrijpend lezen (in engere zin), Opzoeken en Samenvatten - sluiten bij deze interpretatie aan. De correlaties van deze drie verschillende inhoudelijke subvaardigheden variëren van 0.83 tot 0.90 (zie tabel 6.1). Zoals eerder in hoofdstuk 3 is besproken, leggen de toetsen Begrijpend lezen vanaf groep 6 op basis van het referentiekader en in overeenstemming met het onderwijsaanbod in de bovenbouw van het basisonderwijs sterker dan de toetsen voor eerdere leerjaren het accent op studerend lezen. Daarom zijn opgaven toegevoegd over Samenvatten en Opzoeken,

welke dus op één en dezelfde vaardigheidsschaal zijn te brengen als de opgaven over Begrijpend lezen (in engere zin).

Uit het onderzoek naar itemkwaliteit bleek dat de gemiddelde moeilijkheidsgraad van de toets B8/M8 op het vooraf gewenste niveau lag, namelijk tussen 0.65 en 0.75. Daarmee is de toets niet te moeilijk en wordt voorkomen dat de leerling gefrustreerd raakt tijdens de toetsafname ('succeservaring'), waarbij tevens voor een goede spreiding van moeilijkheid over de items gezorgd is. De gemiddelde moeilijkheidsgraad van de toets B8/M8 lag voor afnamemoment B8 op 0.70 en de range van p-waarden tussen 0.50 en 0.89. De gemiddelde moeilijkheidsgraad van de toets B8/M8 lag voor afnamemoment M8 op 0.71 en de range van p-waarden tussen 0.51 en 0.89 (zie tabel 6.2). De gemiddelde Rit-waarde van de toets B8/M8 is voor afnamemoment B8 gelijk aan 0.35 met een range van 0.19 – 0.56 en voor afnamemoment M8 gelijk aan 0.36 met een range van eveneens 0.19 – 0.56 (zie tabel 6.2). De Rit-waarde is op een en hetzelfde item voor B8 en M8 na voor alle items groter dan 0.20 (in het COTAN Beoordelingssysteem de ondergrens voor de beoordeling 'voldoende'). Een Rit-waarde van 0.30 of groter wordt door COTAN als goed gekwalificeerd. Met gemiddelde Rit-waarden van 0.35 en 0.36 is de itemkwaliteit van de toetsen dus goed te noemen.

Uit de resultaten van de kalibratieanalyses viel al af te leiden dat de kwaliteit van de items hoog is, hetgeen dus bevestigd wordt door de 'klassieke' itemparameters: zowel de p-waarden als de Rit-waarden zijn goed te noemen volgens de COTAN-criteria.

Onderzoek naar Differentieel Item Functioneren (DIF) met betrekking tot jongens en meisjes per toetsopgave liet zien dat er bij twee items in lichte mate sprake was van differentieel functioneren. Voor de toets Begrijpend lezen van groep 8 is er dus geen tot nauwelijks sprake van itembias met betrekking tot sekse. Ook de grafische weergaven van de twee items met DIF laten geen bijzonderheden zien (zie figuur 6.1).

Correlaties van de toetsen Begrijpend lezen 3.0 groep 8 met andere LVS-toetsen voor leervorderingen waren conform verwachting. De correlatie met de vaardigheidsscores op basis van LVS 2.0 Begrijpend lezen (een soortgenoot, waarvan de constructvaliditeit positief is beoordeeld), is hoog voor afnamemoment M8. Ook de correlaties tussen toetsscores voor Begrijpend lezen 3.0 op verschillende afnamemomenten (M7, E7 en M8) zijn hoog, d.w.z. de convergente validiteit is hoog. Verder bleek dat de correlaties met andere leervorderingstoetsen die op hetzelfde moment zijn afgenomen, lager zijn dan de correlatie voor afnamemoment M8 met LVS 2.0 M8 (zie tabel 6.5), hetgeen als evidentie voor divergente validiteit kan worden opgevat. Er kan dus geconcludeerd worden dat de convergente en divergente validiteit voldoende was.

Als laatste werden verschillen tussen relevante subgroepen (naar leeftijd, sekse en leerlinggewicht) gepresenteerd. De resultaten bleken aan te sluiten bij de verwachtingen die op grond van theoretische inzichten en eerder onderzoek konden worden geformuleerd. Jongere/versnelde leerlingen (jonger dan 11) doen het aanzienlijk beter dan andere leeftijdsgroepen (effectgrootte = 1.20 t.o.v. het

algemeen gemiddelde) en de oudste groep leerlingen (ouder dan 12, veelal doubleerders) scoort het laagst (effectgrootte = -0.22). Er moet hierbij echter wel opgemerkt worden dat de jongste groep (jonger dan 11) uit slechts 11 leerlingen bestond. De leerlingen die zitten in de jaargroep, die op basis van de leeftijd wordt verwacht, behalen een gemiddelde score die valt tussen de gemiddelde scores van de jongere en oudere leerlingen. Meisjes behalen een iets hogere score dan jongens. Er is sprake van een klein effect (effectgrootte = 0.26). Leerlingen zonder leerlinggewicht (0.00) behalen de hoogste scores (effectgrootte 0.08 t.o.v. het algemeen gemiddelde). Aan de andere kant scoren de leerlingen mét leerlinggewichten (0.30 en 1.20) aanzienlijk lager (effectgrootte -0.42 voor leerlinggewicht 0.30 en effectgrootte -1.53 voor leerlinggewicht 1.20). De leerlingen met ouders met een lagere vooropleiding scoren lager dan leerlingen met ouders met een hogere vooropleiding, waarbij echter wel opgemerkt moet worden dat het aantal leerlingen met een leerlinggewicht anders dan 0.00 slechts zeer gering is.

Al deze resultaten vormen een psychometrische ondersteuning voor de constructvaliditeit van de toetsen. De data geven aan dat er met de toetsen Begrijpend lezen 3.0 gemeten wordt wat men beoogt te meten, namelijk Begrijpend lezen groep 8.

Conclusie:
Op aspect V2 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 op dit aspect het oordeel **'voldoende'** toegekend.

Het volg-aspect

<u>Va1 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:

Uit het kalibratieonderzoek in hoofdstuk 4 blijkt dat de items van de toetsen Begrijpend lezen 3.0 groep 8 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 wordt aan de toetsen Begrijpend lezen groep 8 het oordeel '**voldoende**' toegekend.

<u>Va2 Wordt de betrouwbaarheid van de groei op die schaal adequaat weergegeven?</u>

Bevindingen:
In de WV 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 Begrijpend lezen moet beseft worden dat vaardigheidsgroei zich langzaam in de tijd voltrekt. Dit blijkt ook uit tabel 6.4 van de WV waar een overzicht van de vaardigheidsverdelingen per normeringsmoment (E3 t/m M8) voor Begrijpend lezen wordt gegeven. Tabel 6.4 laat zien dat de gemiddelde vaardigheid van de leerlingen toeneemt van afname tot afname. De gemiddelde vaardigheid op het E-moment is telkens maar iets hoger dan op het M-moment. Dat ondersteunt het advies om de toets Begrijpend lezen slechts één keer per schooljaar af te nemen. In de handleiding bij de toetsen wordt dan ook expliciet vermeld dat het niet nodig is om tweemaal in een schooljaar een toets Begrijpend lezen af te nemen, omdat de groei die leerlingen bij Begrijpend lezen doormaken niet zo groot is.

Conclusie:
Op aspect Va2 wordt voor de toetsen Begrijpend lezen 3.0 groep 8 het oordeel '**voldoende**' toegekend.

Va3 Worden er gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden?

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 rapportagemogelijkheden.

Ook wordt in hoofdstuk 3 van de handleiding ('Nakijken en verwerken van toetsgegevens') aan de hand van een leerlingrapport de interpretatie van groei op een duidelijke manier beschreven. De 67% betrouwbaarheidsintervallen van de vaardigheidsscores worden zowel op het leerlingrapport als in afzonderlijke tabellen vermeld.

Zowel in de WV als de Handleiding wordt het volgen van groei van leerlingen adequaat toegelicht.

Conclusie:
Op aspect Va3 wordt voor de toetsen Begrijpend lezen 3.0 groep 8 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 school periode) 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. Daarbij wordt vooral 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 leerkracht 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. De informatie biedt goede handvatten voor de gesprekken met ouders.

In hoofdstuk 8 van de Handleiding worden veelgestelde vragen (FAQs) behandeld die weliswaar voor de leerkrachten bestemd zijn, maar waar de antwoorden voor een deel ook informatief zijn tijdens bijvoorbeeld de tienminutengesprekken.

In de communicatie naar ouders toe over de resultaten van begrijpend lezen, is het van belang om ook de koppeling te maken naar de resultaten op technisch lezen en woordenschat, omdat die van invloed zijn op de scores van begrijpend lezen.

De taalprofielen 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.

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 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel **'voldoende'** toegekend.

Referentieniveaus

R1 Sluit de inhoud van de toets aan op de kennis en vaardigheden zoals omschreven in de referentieniveaus van het betreffende domein (voor toetsen vanaf groep 6)?

Bevindingen:
Hoofdstukken 2 en 3 van de wetenschappelijke verantwoording geven in samenhang uitgebreid informatie over de operationalisering van de inhoud van de referentieniveaus naar de vorm en inhoud van de toetsmaterialen en soorten opgaven. Uit deze teksten blijkt de jarenlange ervaring van de toetsaanbieder met constructie van toetsen op het gebied van begrijpend lezen, en van omgang met de beschrijvingen in het Referentiekader Taal.

Conclusie:
Op aspect R1 wordt aan de toetsen Begrijpend lezen 3.0 groep 8 het oordeel **'voldoende'** toegekend.

3. Verzamelstaat

Kwaliteitsaspect	Code	Oordeel
De kwaliteit van de steekproef	*S1*	**Voldoende**
	S2	**Voldoende**
Normering	*N1.1*	**n.v.t.**
	N1.2	**n.v.t.**
	N1.3	**n.v.t.**
	N2.1	**Voldoende**
	N2.2	**Voldoende**
Betrouwbaarheid	*B1*	**Voldoende**
	B2	**Voldoende**
Validiteit	*V1*	**Voldoende**
	V2	**Voldoende**
Volg-aspect	*Va1*	**Voldoende**
	Va2	**Voldoende**
	Va3	**Voldoende**
Inzicht in leervorderingen	*I1*	**Voldoende**
Referentieniveaus	*R1*	**Voldoende**

4. Literatuurlijst

- Tomesen, M., Engelen, R. en Hiddink, L. (2019). *Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 8*. Arnhem: Cito B.V.
- Cito (2018). *Cito Volgsysteem primair en speciaal onderwijs. Begrijpend lezen 3.0 Groep 8*. Arnhem: Cito B.V.