

## **CITO LOVS WOORDENSCHAT GROEP 8**

Deze beoordeling betreft zowel de papieren als de digitale versie van de toets.

### ***Theoretische Uitgangspunten***

De toets Cito LOVS Woordenschat Groep 8 beoogt een beeld te geven van zowel de ontwikkeling van de breedte (de omvang) van de woordenschat als ook de diepte ervan (de relatie van woorden en begrippen tot andere woorden en begrippen) in het Nederlandse primair onderwijs groep 8. Wat het construct ‘woordenschat’ inhoudt, wordt uitvoerig theoretisch ingekaderd, zij het niet zozeer in de context van intelligentietheorieën als in het kader van de cognitieve ontwikkeling van kinderen. Uitvoerig wordt verantwoord welke inhoud de auteurs in de toets willen onderbrengen en op welke aspecten van de woordenschat deze een beroep doet. Op basis van een zorgvuldige analyse zijn de items inhoudelijk verdeeld naar betekenis en betekenisrelaties van woorden, en naar inhoudswoorden en functiewoorden. Het is echter niet duidelijk welke criteria of theorie de auteurs hebben gebruikt om de subcategorieën te definiëren. Hoewel zij stellen dat er geen detailanalyse op deze categorieën of subcategorieën beoogd wordt, zou het toch wenselijk zijn als de operationalisering van de categorieën meer zou worden toegelicht.

De auteurs geven aan dat woordenschat een multidimensioneel karakter heeft waarbij het kind een woord op verschillende niveaus kan beheersen. Hiermee wordt ook rekening gehouden bij de constructie van de items. Het is jammer dat de uiteindelijke score van het kind slechts een totaalscore is waarbij dus niet gedifferentieerd wordt naar de (kwalitatief) verschillende niveaus van woordenschat. Vanuit praktisch oogpunt is het weliswaar wenselijk om tot één globale totaalscore te komen voor ieder kind, maar vanuit diagnostisch oogpunt zou het zeer informatief zijn wanneer de test een profiel zou geven van de onderscheiden woordenschatniveaus van het kind.

Naast niveaubepaling en progressiebepaling kan de toets worden gebruikt voor probleemsignalering. Deze signaleringsfunctie is echter nog niet uitgewerkt. Meer wetenschappelijk onderzoek is nodig om de foutenanalyse en mogelijkheden tot remediëring verder te onderzoeken.

De toets is bedoeld voor leerlingen uit het regulier basisonderwijs in groep 8, maar is ook toepasbaar bij leerlingen op speciale scholen voor basisonderwijs en voor ‘speciale’ leerlingen in het reguliere onderwijs. Nog maar kort in Nederland verblijvende kinderen zijn uitgesloten. Verder wordt duidelijk gemaakt dat, gelet op de normering, de geëigende afnamemomenten november (B8) en januari (M8) zijn.

### ***Kwaliteit van het Testmateriaal***

#### Papier-en-potloodversie

De opgaven zijn gestandaardiseerd, correct geformuleerd en vrij van kwetsende inhoud. In de stam van een aantal items komt het woord ‘qua’ voor. Het is de vraag of ‘qua’ tot de woordenschat van groep 8 behoort. Een alternatief zou zijn om ‘wat de... betreft’ te gebruiken in plaats van ‘qua’.

Bij de scoring, die aan de hand van een lijst met goede antwoorden gaat (dus niet met een scoringsmal die over het antwoordformulier wordt gelegd), wordt een beroep gedaan op de nauwkeurigheid van de testleider. Een controlemogelijkheid is er niet.

De instructie voor de geteste is volledig en duidelijk. Tevens wordt een ruim aantal voorbeelden aangeboden. Hoewel de pagina-indeling van de boekjes op de antwoordformulieren niet terugkomt als blokken, zijn de antwoordformulieren overzichtelijk genoeg. Het is wel de vraag of het werken met een apart antwoordformulier niet gevoelig is voor fouten, omdat leerlingen redelijk gemakkelijk een antwoord van de verkeerde opgave kunnen omcirkelen.

Al het testmateriaal ziet er verzorgd en aantrekkelijk uit.

#### Computerversie

De items van de digitale versie van de toets zijn voor een groot deel dezelfde als die van de papieren versie. De mix van de opgavetypes tussen de papieren versie en de digitale versie is miniem anders. Dit wordt weergegeven in de tabellen 3.4 en 3.6 in de Wetenschappelijke Verantwoording. De digitale versie wordt automatisch gescoord, waarbij het vlot en zoveel mogelijk profiterend van alle mogelijkheden van de invoer, analyse en interpretatie, grondige kennis van het programma LOVS vereist. Overigens worden de scores op

de digitale versie niet hetzelfde verwerkt als de scores op de papieren versie. Bij de digitale versie worden de items gewogen met een discriminatie-index. Op die manier kan men zowel een ongewogen als gewogen score berekenen. Het voordeel van de gewogen score is dat de standaardfout van de schatting kleiner is. Het toepassen van de discriminatie-index kan bij de digitale versie automatisch worden gedaan. Bij de papieren versie moet dit handmatig gedaan worden voor het gehele antwoordpatroon, wat veel tijd vergt van de leerkrachten. Overigens heeft het Cito in aanvullende informatie aan de COTAN laten weten dat de correlaties tussen de schattingen van de gewogen en ongewogen scores voor de verschillende afnamemomenten alle groter dan .99 zijn. Bovendien laten de resultaten van verschillende t-toetsen zien dat er voor geen van de leerlingen een significant verschil aanwezig is tussen de gewogen en ongewogen score. De instructie voor de geteste is precies en compleet. De gebruikersinterface is een goede mix van functionaliteit en aantrekkelijkheid.

Het is als gebruiker lastig een volledig beeld over de beveiliging van het Cito Portal te vormen, maar aan basale eisen (een beveiligde internetomgeving en een aanmeldingsprocedure met wachtwoord) is voldaan.

### ***Kwaliteit van de Handleiding***

In de handleiding wordt niet alleen over de gewone en bijzondere gebruiksmogelijkheden informatie gegeven (bijvoorbeeld wat te doen bij een bijzonder slecht resultaat), maar ook over beperkingen (bijvoorbeeld het gebruik van de toets bij kinderen die nog maar kort in Nederland verblijven). Er ontbreekt alleen nog het beloofde 'Hulpboek Woordenschat Groep 8', zoals genoemd in paragraaf 4.3 van de Handleiding. Het is aan te raden dit hulpboek bij een volgende beoordeling aan te bieden, want dan kan er iets meer gezegd worden over de hulpverlening aan kinderen met een ontoereikende woordenschat. De wetenschappelijke onderbouwing van signaleren, analyseren en handelen ontbreekt.

Er staat in de handleiding geen samenvatting van de onderzoeksgegevens. De onderzoeksgegevens worden alleen in de Wetenschappelijke Verantwoording verstrekt. Het is niet duidelijk of het beleid bij Cito is dat de gebruiker/diagnosticus deze verantwoording standaard verkrijgt. Bovendien is die moeilijk voor een niet-specialist en dat is een reden te meer om essentiële zaken (bijvoorbeeld over betrouwbaarheid) op een toegankelijke manier in de gebruikershandleiding uit te leggen.

De aanwijzingen voor de testleider zijn volledig en uitgebreid en maken voor de testleider duidelijk wat hij/zij moet en mag zeggen voor en tijdens de klassikale afname. Er wordt ook goed uitgelegd dat de meer afstandelijke rol van een testleider een andere is dan die van een leerkracht, die van nature gericht is op het geven van hulp. Leerkrachten basisonderwijs zijn de beoogde testleiders. Het is de vraag of voor het analyseren van gegevens op groeps- en/of schoolniveau en het vertalen daarvan naar kwaliteitsbeleid niet enige extra training vereist zou moeten worden.

Er wordt één beknopt voorbeeld van een leerlingenrapport gegeven, dat echter niet in de context wordt geplaatst van andere gegevens over de ontwikkeling van het kind. Er is enigszins aandacht voor andere soorten informatie die van belang zijn bij de interpretatie.

De aanmeldingsprocedure bij de Cito Portal en de daarop volgende installatiestappen worden duidelijk uitgelegd, net als informatie over de bediening en mogelijkheden ervan.

De helpdesk is telefonisch ruim bereikbaar op werkdagen en ook de responstijd op een via de mail gerapporteerd probleem is prima.

### ***Normen***

De normgegevens zijn verzameld in november 2011 (B8) en januari 2012 (M8). De normgroepen zijn met 1551 leerlingen voor afnamemoment B8 en 755 leerlingen voor afnamemoment M8 qua grootte 'goed' te noemen. Voor het onderzoeken van de representativiteit van de normgroepen zijn verschillende achtergrondvariabelen gebruikt zoals percentage achterstandsleerlingen, geografische spreiding en urbanisatie. Informatie over de representativiteit ontbreekt op leerlingniveau, maar aangenomen mag worden dat de verhouding jongens/meisjes en de informatie over de leeftijd overeenkomen met die in de populatie door het betrekken van complete jaargroepen in de normering. In feite fungeert dus niet de leerling, maar in eerste instantie de school, als steekproefeenheid, hetgeen uitgebreid wordt verantwoord. De representativiteit op de genoemde criteria laat geen afwijkingen van betekenis zien.

*Documentatie van Tests en Testresearch in Nederland*  
**TOELICHTING BIJ DE BEOORDELING**

De normscores kunnen worden uitgedrukt in de twee van het Cito bekende normschalen, namelijk de uniforme verdeling in de klassen I t/m V en de verdeling in A (25%), B (25%), C (25%), D (15%) en E (10%). Deze normschalen komen goed overeen met het doel van de toets en worden duidelijk uitgelegd. Het zou voor een leerkracht onduidelijk kunnen zijn waarom twee niveau-indelingen (Romeinse cijfers en letters) worden gepresenteerd en op welke hij/zij zich moet richten. In de handleiding wordt vrijwel in het midden gelaten welke niveau-indeling de leerkracht zou kunnen gebruiken. Er is een lichte voorkeur voor I-V, maar dit zou meer toegelicht kunnen worden aangezien een leerkracht niet deskundig hoeft te zijn op het gebied van de statistische kenmerken van verdelingen van (norm)scores.

Gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling worden vermeld. De lokale meetnauwkeurigheid van de toets wordt grafisch weergegeven. Hieruit blijkt dat de toets telkens het meest nauwkeurig is in de lagere en de gemiddelde vaardigheidsniveaus. De betrouwbaarheid is onderzocht door middel van de coëfficiënt MAcc voor zowel de papieren als de digitale versie (bij de laatste is de score de som van de met de discriminatie-index gewogen itemscores). Uit de MAcc is vervolgens de standaardmeetfout berekend.

Mogelijke verschillen tussen de vier subgroepen die gevormd worden door de vier strata en verschillen tussen de subgroepen naar regio (voor zowel B8 als M8) zijn onderzocht en niet gevonden. Doordat er geen informatie verzameld is op het leerlingniveau, is het verschil tussen autochtone en allochtone leerlingen niet onderzocht. Daarnaast zou onderzoek naar verschillen in scores tussen jongens en meisjes aanvullende informatie hebben kunnen opleveren, bijvoorbeeld naar de eventuele wenselijkheid van aparte normering. Het is niet duidelijk waarom de sekseverschillen niet naderhand aan de hand van automatisch geretourneerde data zijn bekeken.

NB. De normering van de papieren toetsen is tevens toepasbaar op de digitale toetsen. Dit omdat het kalibratieonderzoek in paragraaf 4.1 van de Wetenschappelijke Verantwoording laat zien dat de digitale en papieren opgaven op dezelfde vaardigheidsschaal passen. Daarnaast heeft het Cito in een reactie aan de COTAN aanvullende ondersteunende informatie gestuurd. Hieruit blijkt dat de latente correlatie tussen de geschatte vaardigheid op basis van papieren opgaven en de geschatte vaardigheid op basis van digitale opgaven voor B8 .94 is. Tevens blijkt dat voor 98.1% van de leerlingen die twee schattingen niet significant van elkaar verschillen op afnamemoment B8.

Overigens ontbreken bij Figuur 4.1 de namen van de  $x$ -as en  $y$ -as en is het daardoor niet duidelijk welke informatie in deze figuur wordt weergegeven. Het Cito heeft in een reactie laten weten dat de vaardigheid is weergegeven op de  $x$ -as en de score op de toets op de  $y$ -as. Voor deze figuur is zowel voor de papieren als voor de digitale toets de ongewogen score gebruikt.

### ***Betrouwbaarheid***

In de Wetenschappelijke Verantwoording wordt informatie aangedragen over de betrouwbaarheid. De MAcc waarden zijn .90 (B8 en M8) voor de papieren versie en .91 (B8 en M8) voor de digitale versie. Daarmee zijn deze waarden 'goed' voor belangrijke beslissingen. Inzicht in de lokale betrouwbaarheid wordt ook verschaft door verwarringsmatrices tussen de (A t/m E) geschatte en werkelijke vaardigheidsscores.

Test-hertestbetrouwbaarheden zijn berekend op basis van een simulatieonderzoek ( $n = 100.000$ ). Bij een bepaalde aselechte vaardigheid uit de totale vaardigheidsverdeling van alle leerlingen zijn steeds twee onafhankelijke, bij de geselecteerde vaardigheid horende toetsafnames, gesimuleerd. Voor beide afnamemomenten en beide versies (papier en digitaal) zijn de uitkomsten identiek aan de hierboven genoemde MAcc-coëfficiënten.

### **Begripsvaliditeit**

De basis voor de begripsvaliditeit wordt gelegd door het feit dat de items uit een OPLM-gekalibreerde itembank afkomstig zijn (waarin zij passen) en dus een beroep doen op hetzelfde complex aan vaardigheden. Echter omdat de toets ééndimensionaal gemaakt is door items te verwijderen die niet in het meetmodel pasten, is het een cirkelredenering om te concluderen dat de constructvaliditeit in orde is *omdat* de items een ééndimensionale schaal vormen. Wanneer een test op deze wijze ééndimensionaal is gemaakt, kan men op basis van de dimensionaliteit geen uitspraken meer doen over de constructvaliditeit. Andere aanwijzingen zijn nodig. Gegevens die daartoe bijdragen worden verstrekt in het overzicht van de item gemiddelden, de laagste en de hoogste *p*-waarde, de item-testcorrelaties en de item-restcorrelaties. Slechts één opgave in de toets van 70 opgaven (het is niet bekend of het gaat om de B8 of M8 toets) heeft een item-restcorrelatiekleiner dan .20.

Verder is de normscore gecorreleerd met (onderdelen van) de Eindtoets Basisonderwijs ter ondersteuning van de convergente en divergente validiteit. De correlaties laten zien dat de scores op de Woordenschat Groep 8 sterk samenhangen met het onderdeel Taal (minus Woordenschat) van de Eindtoets ( $r = .82$ ) en minder sterk met het onderdeel Rekenen-Wiskunde ( $r = .59$ ), waarbij de correlatie met Rekenen-Wiskunde wel hoger is dan verwacht zou kunnen worden. Overigens is de correlatie met het onderdeel studievaardigheden .79; daarop wordt niet ingegaan. Hetzelfde geldt voor de toch wat tegenvallende correlatie met het onderdeel Spelling ( $r = .52$ ). Bij dit onderzoek ontbreken de gegevens over de gebruikte steekproeven. Hierbij dient opgemerkt te worden dat ervan uit is gegaan dat deze gegevens alleen betrekking hebben op de toets voor Groep 8, dit staat overigens nergens expliciet vermeld.

Om te beoordelen of de Woordenschat Groep 8 daadwerkelijk woordenschat meet zouden de auteurs een andere woordenschat test moeten gebruiken (bijvoorbeeld de WAT of eventueel de TAK), vergelijking met ander materiaal dan de eigen Cito gegevens had voor de hand gelegen.

De samenhang met leerjaar wordt aangetoond door de vaardigheidsverdelingen van B8 en M8 te vergelijken, waarbij een toename te zien is, die ook verwacht werd op basis van leeftijd, ontwikkeling en genoten onderwijs, over deze (toch wel) korte periode tussen begin (oktober-november) en midden (januari-februari) groep 8.

NB. Het kalibratie-onderzoek is zodanig verricht dat de papieren en digitale versie als equivalent kunnen worden beschouwd, zodat de gegevens over de begripsvaliditeit op beide versies van toepassing zijn. Zo verschillen bijvoorbeeld de *p*-waarden van de digitale versie en de papieren versie nauwelijks van elkaar (zie tabel 4.7 van de Wetenschappelijke Verantwoording). Daarnaast zijn de items van de digitale versie van de toets voor een groot deel dezelfde als die van de papieren versie. Daarbij is de mix van de opgaventypes tussen de papieren versie en de digitale versie miniem anders. Dit wordt weergegeven in de tabellen 3.4 en 3.6 in de Wetenschappelijke Verantwoording.

### **Criterionvaliditeit**

Deze test is volgens de auteur(s)/uitgever niet bedoeld voor voorspellend gebruik, criteriumvaliditeit is derhalve niet van toepassing. Wanneer de test wordt ingezet in situaties waarin voorspelling wel aan de orde is, geldt het oordeel onvoldoende, omdat er geen onderzoek naar de criteriumvaliditeit is verricht.