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 Begrijpend Luisteren groep 4 zijn een onderdeel van het Cito Volgsysteem primair onderwijs (tweede generatie toetsen) en zijn bedoeld voor leerlingen in groep 4 van het primair onderwijs. De toets vormt samen met de LVS toetsen Begrijpend Luisteren voor groep 3 tot en met 8 een systeem om vast te stellen hoe leerlingen met begrip kunnen luisteren en hoe hun luistervaardigheid zich in de basisschoolperiode ontwikkelt. Onderstaande beschrijving is gebaseerd op de Handleiding en de Wetenschappelijke verantwoording.

Meetpretentie

De toetsen in de toets pakketten Begrijpend Luisteren groep 4 van het Cito Volgsysteem primair en speciaal onderwijs zijn bedoeld om vast te stellen hoe goed een leerling een betekenis kan toekennen aan gesproken taal en of het adequaat kan reageren op gesproken taal en hoe de vaardigheid in begrijpend luisteren van de leerling zich in de loop van de jaren ontwikkelt.

Doelgroep

De toetsen Begrijpend Luisteren groep 4 zijn bedoeld voor leerlingen in groep 4 van het primair en speciaal onderwijs, maar kunnen ook gebruikt worden voor leerlingen uit andere jaargroepen die werken op het niveau van groep 4 en voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften. Voor deze groepen speciale leerlingen zijn geen afzonderlijke normen vastgesteld en de toetsresultaten van deze leerlingen worden geïnterpreteerd met behulp van de gemiddelde vaardigheidsscores voor leerlingen uit het regulier onderwijs.
Voor leerlingen die nog maar pas in Nederland verblijven of gehoorproblemen hebben, zijn de toetsen ongeschikt. Een leerling moet voldoende taalvaardig in het Nederlands zijn (vergelijkbaar met Nederlandstalige leerlingen in groep 4) voordat de toets Begrijpend Luisteren kan worden afgenomen.

Gebruiksdoel en functie

Doel van de toetsen Begrijpend Luisteren groep 4 is het in kaart brengen van het vaardigheidsniveau en de ontwikkeling van de leerlingen op het gebied van begrijpend luisteren. Hiervoor wordt de behaalde vaardigheidsscore normgericht geïnterpreteerd op basis van de vaardigheidsverdeling in een adequate, landelijke, referentiegroep. De vaardigheidsscore wordt uitgedrukt in de symmetrische niveau indeling I t/m V en in de asymmetrische niveau indeling A t/m E. De toetsen maken het mogelijk om:

- De vaardigheid begrijpend lezen 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 begrijpend lezen 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 toets Begrijpend Luisteren is gebaseerd op de door het OCW geformuleerd kerndoelen en de door het project TULE ontwikkelde tussendoelen en leerlijnen. De kerndoelen voor het onderwijs in begrijpend luisteren zijn grotendeels ondergebracht bij 'Mondeling taalonderwijs', kerndoel 1 en bij 'Taalbeschouwing', kerndoel 12. In het project TULE zijn de kerndoelen uitgewerkt in tussendoelen en leerlijnen per basisschoolgroep. Het construct luisteren wordt hierbij gedefinieerd als de interactie tussen begrijpen, interpreteren en reflecteren wat plaatsvindt binnen een context die steeds vaker wordt gevoed met zowel auditieve als visuele stimuli ('de beeldcultuur').

Inhoud van het toetspakket
Het toetspakket Begrijpend Luisteren groep 4 bestaat uit de volgende documenten:
- Handleiding, deze bevat informatie over:
  - de afname van de toets (hfdst. 2),
  - nakijken en verwerken van toetsgegevens (hdfst. 3),
  - interpretatie van de toetsresultaten op leerling- en groepsniveaus (hfdst 4),
  - interpretatie van toetsresultaten op schoolniveau (hfdst 5),
  - theoretisch kader en achtergronden van de toets (hfdst 6),
  - communiceren over toetsresultaten met leerling en ouders (hfdst 7),
  - achtergrondinformatie en veelgestelde vragen (hfdst 8) en
  - enkele bijlagen
- Één toets:
  - Toets M4/E4
- Afnamekaarten met aanwijzingen voor de papieren of de digitale afname van de toetsen
- Nakijkkaarten
- Antwoordbladen
- Tabellen voor de toets om de vaardigheidsscore en –niveau te bepalen.

2. Beoordeling van de kwaliteitsaspecten

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

De kwaliteit van de steekproef

S1.1. Is de steekproef representatief?

Bevindingen:
In november 2010 zijn in een kalibratieonderzoek (proefonderzoek) 164 items voorgelegd aan 757 leerlingen van groep 4. De 164 items waren verdeeld over 8 boekjes (booklets).

Elk boekje bestond uit circa 40 opgaven. Elke opgave kwam in twee boekjes voor. Het gemiddeld aantal antwoorden per item was 186 wat boven het minimum vereiste van 150 ligt.

Op grond van het kalibratieonderzoek (nov. 2010) is een selectie van items gemaakt voor de normeringsonderzoeken M4 en E4. De normeringen vonden plaats in respectievelijk jan/feb 2012 (M4) en mei/juni 2012 (E4). Voor de normeringsonderzoeken werd gebruik gemaakt van 1.785 leerlingen, afkomstig van 76 scholen (M4) en van 1.961 leerlingen afkomstig van 76 scholen (E4).

De representativiteit van de steekproeven voor de normeringsonderzoeken M4 en E4 is onderzocht met betrekking tot regio, urbanisatiegraad, schooltype en geslacht.

Bij regio is uitgegaan van de vier landsdelen / regio's van de CBS-indeling. In de steekproef van normeringsmoment M4 zijn de scholen vrijwel exact verdeeld zoals de landelijke verdeling van scholen. In de steekproef van normeringsmoment E4 is de regio West wat ondervertegenwoordigd en de regio Zuid wat oververtegenwoordigd. Om te bepalen of een weging noodzakelijk is, zijn de gemiddelden en standaarddeviaties van de steekproef E4 berekend. De gevonden effectgroottes waren zodanig klein dat weging niet noodzakelijk was.

Bij urbanisatiegraad is uitgegaan van de CBS-indeling naar vijf niveaus van mate van verstedelijking. In de steekproef M4 zijn de scholen in weinig verstedelijkte gebieden wat oververtegenwoordigd en de scholen in niet verstedelijkte gebieden wat ondervertegenwoordigd. In de steekproef E4 zijn de scholen in zeer sterk verstedelijkte gebieden licht ondervertegenwoordigd en scholen in niet verstedelijkte gebieden wat oververtegenwoordigd. De gevonden verschillen bleken niet significant.

Bij schooltype is gebruik gemaakt van de formatiegewichten van leerlingen volgens de meest recente regeling van OCW. Op basis van gewicht wordt er een indeling gemaakt naar vier groepen scholen. En op basis van schoolgrootte wordt er een indeling gemaakt in twee subgroepen. In totaal ontstaan er zo acht strata. In de steekproef M4 is stratum 2 licht ondervertegenwoordigd en zijn stratum 5 en 7 heel licht oververtegenwoordigd. In de steekproef E4 is stratum 2 juist licht oververtegenwoordigd evenals stratum 1 en is stratum 4 licht ondervertegenwoordigd. De gevonden verschillen bleken niet significant.

Wat betreft geslacht zijn de leerlingen in de steekproef van normeringsmoment M4 vrijwel exact verdeeld zoals de landelijke verdeling van leerlingen. Ook voor het normeringsmoment E4 zijn de verschillen minimaal en niet significant.

Conclusie:
De steekproef is representatief tav vier relevante variabelen en aan aspect S1.1. kan dan ook het oordeel '**voldoende**' toegekend worden.

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 het proefonderzoek van november 2010 en de opgaven uit de twee normeringsonderzoeken in 2012 gekalibreerd ('geijkt'). Hiervoor is gebruik gemaakt van het IRT model OPLM (One Parameter Logistic Model). Met dit statistische model zijn de

psychometrische kenmerken (moeilijkheidsparameters en discriminatie indices) van de items geschat.

Kenmerkend voor een IRT model is dat, in tegenstelling tot het KTT model, het te meten begrip (latente vaardigheid) centraal staat. Binnen de IRT is de moeilijkheid van een item gedefinieerd in termen van de onderliggende vaardigheid, zonder enige verwijzing naar een bepaalde populatie van leerlingen (wat juist kenmerkend is voor de KTT). Binnen de IRT liggen moeilijkheid en vaardigheid dan ook op hetzelfde unidimensionele continuüm.

In OPLM worden de eigenschappen van het Rasch model gecombineerd met de flexibiliteit van het twee-parameter logistisch model. Zo worden itemmoeilijkheidsparameters geschat en itemdiscriminatieparameters geïmputeerd als bekende constanten. Tevens wordt nagegaan of de opgaven van een onderdeel kunnen worden beschreven met een unidimensionele onderliggende vaardigheid. Dit alles vindt plaats via een iteratief proces waarin alternerend de modelfit van items wordt onderzocht.

Het algemene (inhoudelijke) uitgangspunt voor de toets Begrijpend Luisteren is dat de (latente) vaardigheid Begrijpend Luisteren kan worden opgevat als een unidimensionaal continuüm en dat elke leerling kan worden voorgesteld als een punt op dit continuüm, rekening houdend met een zekere mate van onnauwkeurigheid (vergelijkbaar met de standaardmeetfout uit de KTT). Hierbij wordt opgemerkt dat de elementen Begrijpen, Interpreteren en Reflecteren als de drie componenten van de latente vaardigheid Begrijpend Luisteren niet opgevat kunnen worden als te isoleren vaardigheden. Daarmee wordt de vaardigheid Begrijpend Luisteren als interactie tussen deze drie componenten, beschouwd als één unidimensionele vaardigheid.

Met kalibratie wordt bedoeld dat er kengetallen worden gezocht bij de items die de antwoorden van de leerlingen goed representeren. In OPLM worden met de CML-methode (Conditional Maximum Likelihood) de itemparameters geschat en wordt gecontroleerd of deze de data goed voorspellen. Na het schatten van de itemparameters met behulp van de CML-methode wordt met behulp van de M-toetsen gecontroleerd of de discriminatie indices goed zijn ingesteld. Hierna volgt onderzoek naar de passing van de items met behulp van de S-toetsen. Tot slot vindt een globale modelcontrole plaats. Uit de visuele beoordeling van de S-toetsen volgt een zeer sterke aanwijzing dat het meetinstrument en het meetmodel adequaat het gedrag van de leerlingen verklaren. Zowel items als de gehele toets passen bij het model.

In het kalibratieproces is uitgegaan van een onvolledig maar ‘verbonden’ design. In het kalibratieonderzoek van november 2010 zijn 164 items voorgelegd aan 757 leerlingen van groep 4. De 164 items waren verdeeld over 8 boekjes (booklets). Elk boekje bestond uit circa 40 opgaven, bestaande uit ongeveer 20 items van twee taken M4E4. Elke opgave kwam in twee boekjes voor. Het gemiddeld aantal antwoorden per item was 186 wat boven het minimum vereiste van 150 ligt.

Daarmee is bij het proefonderzoek sprake van een onvolledig design (ankeritem design) dat is verbonden door ankers die uit circa 20 items bestaan. De 20 items van taak 8 vormen de link met de toets Begrijpend Luisteren groep 3.

Op basis van inhoudelijke en psychometrische criteria werden 32 items voor de toets M4/E4 geselecteerd. De 32 vragen zijn verdeeld naar vaardigheid en inhoudsaspecten van

de componenten Begrijpen en Interpreteren. Van alle opgaven die zijn meegegaan in het normeringsonderzoek zijn de gekalibreerde p-waarde, de rit waarde en de rir waarde bepaald. Voor de normeringsonderzoeken M4 en E4 werden scholen geworven na het trekken van een representatieve steekproef waarbij rekening werd gehouden met verdeling over strata, regio en urbanisatiegraad. Voor de normeringsonderzoeken werd gebruik gemaakt van 1.785 leerlingen, afkomstig van 76 scholen (M4) en van 1.961 leerlingen afkomstig van 76 scholen (E4).

Uit het kalibratieonderzoek blijkt dat voor beide afnamemomenten de items passen bij voornoemd IRT model en dat het model ook past voor de toets M4/E4 als geheel. Dit betekent dat er sprake is van één unidimensionele vaardigheidsschaal waar items en leerlingen op afgebeeld kunnen worden.

Conclusie:
Het onvolledige maar 'verbonden' design van het proefonderzoek is adequaat. Het volledige design van toets M4/E4 is 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 twee afnamemomenten in het jaar, het M moment (halverwege het schooljaar) en het E moment (aan het eind van het schooljaar). De toets kent derhalve twee afnamemomenten, maar is dezelfde toets.

Op grond van het kalibratieonderzoek (nov. 2010) is een selectie van items gemaakt voor de normeringsonderzoeken M4 en E4. Voor de normeringsonderzoeken werd gebruik gemaakt van aselecte steekproeven van leerlingen uit de populaties die in overeenstemming zijn met de afnameperiodes M4 en E4. Voor M4 werden toetsresultaten gebruikt van 1.785 leerlingen, afkomstig van 76 scholen. Voor E4 werden toetsresultaten gebruikt van 1.961 leerlingen afkomstig van 76 scholen. De normeringen vonden plaats in respectievelijk jan/feb 2012 (M4) en mei/juni 2012 (E4). Bij de toets werden relatieve normen opgesteld.

Voor beide momenten (M4 en E4) worden vaardigheidsverdelingen gepresenteerd. 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 asymmetrische niveau indeling A t/m E en de symmetrische niveau indeling I t/m V) worden bepaald.

Het blijkt dat de gemiddelde vaardigheid in Begrijpend Luisteren in de periode tussen de afnamemomenten toeneemt, terwijl de spreiding nagenoeg gelijk blijft. De R0 toets impliceert dat de vaardigheden op de normeringsmomenten als normaal verdeeld kunnen worden opgevat.

De normen voor de toets Begrijpend luisteren groep 4 zijn geldig tot en met 2022.

Conclusie:

Aan aspect N1.2.1 wordt het oordeel '**voldoende**' toegekend.

N1.2.2. Zijn de normgroepen representatief?

Bevindingen:
De representativiteit van de steekproeven voor M4 en E4 zijn besproken bij punt S1.1. Daar werd geconstateerd dat de steekproeven voor zowel M4 als E4 representatief zijn tav vier relevante variabelen.

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 KTT. Deze coëfficiënt wordt in de psychometrische literatuur beschreven en als correct aangemerkt.

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:
Voor M4 en E4 worden drie betrouwbaarheidsgegevens berekend: standaardmeetfout, MAcc en een gesimuleerde test-hertest betrouwbaarheidscoëfficiënt. Voor M4 zijn de gegevens gelijk aan 2,44, 0,73 en 0,73 en voor E4 gelijk aan 2,28, 0,72 en 0,72. De auteurs van de wetenschappelijke verantwoording verwijzen naar het beoordelingssysteem van de COTAN waar voor tests die geen zware consequenties voor leerlingen hebben, zoals de toets Begrijpend luisteren volgens de auteurs, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 als 'voldoende' aangemerkt wordt.

Naast klassieke betrouwbaarheidscoëfficiënten is ook de lokale betrouwbaarheid en de meetnauwkeurigheid onderzocht. De betekenis van de meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit de betrouwbaarheidstabellen van twee niveauverdelingen (I t/m V en A t/m E) voor M4 en E4. Uitgaande van de betrouwbaarheidstabellen worden twee indices voor de nauwkeurigheid van de classificaties gerapporteerd: de plus/minus 1 niveau-index en de marginal classification index berekend. Uit de hoogte van de indices blijkt dat de laagst en hoogst scorende leerlingen accuraat te classificeren zijn maar dat tussen leerlinge in de niveaugroepen B, C en D , respectievelijk II, III en IV minder duidelijk onderscheid te maken is.

Conclusie:

De betrouwbaarheid van de toets Begrijpend luisteren 4 is 'voldoende' als aangenomen mag worden dat de toets geen zware consequenties voor de leerlingen heeft en ingestemd wordt met de beoordelingscriteria voor de betrouwbaarheid van de COTAN.

Op aspect B1.2 wordt aan de toets Begrijpend luisteren groep 4 het oordeel '**voldoende**' toegekend.

Validiteit

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

Bevindingen:
Opgaven worden als volgt verdeeld naar vaardigheid en inhoudsaspecten:

Begrijpen van gesproken taal:

- Betekenis woord en woordgroep
- Specifieke inhoudselementen die expliciet in de tekst aan de orde zijn (o.a. meningen, hoofdgedachte, tijdsperioden, handelingen)
- Eenvoudige expliciete verbanden (o.a. vergelijkingen, tegenstellingen)
- Complexe expliciete verbanden (o.a. oorzaak-vervolg, deel-geheel, hoofd en bijzaken, opeenvolgende stappen)

Opgaven bij interpreteren:

- Afleiden van de betekenis van het woord
- Afleiden van informatie uit de tekst (o.a. ontbrekende info, mening, functionele betekenis)
- Globale inhoud van de tekst (o.a. verbinden, en vergelijken van informatie, dus: hoofdgedachte, hoofdpersoon, doel en publiek herkennen, samenvatten)
- Manier van spreken (klemtoon, tempo, toon)

In groep 4 zijn er meer opgaven rond begrijpen (63%). In schema's wordt het aantal vragen weergegeven per inhoudsaspect voor begrijpen en interpreteren. In de handleiding is geen overzicht gegeven bij welke tekst en bijbehorende vragen nu wat getoetst wordt. Dit moet de leraar zelf gaan herleiden. Dit is niet handig zeker als de toetsen ook handvatten moeten geven waar kinderen op uitvallen en waar extra aandacht aan zou moeten worden besteed.

Er wordt gebruik gemaakt van twee teksttypen: fictie en non-fictie; tekstgenres (soms in combinatie ingezet): verhaal (i.c. fragmenten uit speelfilms), nieuwsbericht, documentaire, instructie.

De teksten zijn kort (max. 6 minuten) zodat kinderen hun aandacht kunnen bewaren en hebben een duidelijke opbouw en structuur. De teksten sluiten aan bij de leefwereld van de kinderen (bijv. logeren, neushoorn, paarden verzorgen, kinderlied)

Uitgangspunten bij de toetsconstructie waren met name de kerndoelen primair onderwijs en de tussendoelen mondelinge communicatie. Zowel bij de kerndoelen als de tussendoelen is begrijpend luisteren één van de aspecten van mondeling taalonderwijs. Je toetst hierdoor maar één onderdeel namelijk dat leerlingen leren informatie te verwerven uit gesproken taal. Begrijpend luisteren is één van de drie kerndoelen rond mondelinge

taal. Maar het leren argumenteren of je leren uit te drukken (de andere twee kerndoelen) worden dus niet getoetst. Van de tussendoelen zijn vooral de doelen rond het begrijpen van het verhaal, informatie uit het verhaal halen, afleiden hoofdgedachte, gebruik maken van de structuur van het verhaal herkenbaar in de toetsen en niet het samenvatten van het verhaal in eigen woorden, bewust zijn van de impact van media etc. Opvallend is dat er in de handleiding bij de theoretische inkadering wel kort aandacht wordt besteed aan de uitwerking van de expertgroep Doorlopende Leerlijnen, maar dat de tussendoelen en leerstoflijnen uitgangspunt zijn geweest bij de opzet en uitwerking van de toetsen. Omdat er toegewerkt wordt naar tenminste 1F niveau zou het een meerwaarde hebben als er ook aandacht wordt besteed aan de referentieniveaus. Het domein Mondelinge taalvaardigheid bij de referentieniveaus heeft een niveaubeschrijving uitgewerkt rond luistervaardigheid waarin bij 1F wordt aangegeven dat een kind kan luisteren naar eenvoudige teksten over alledaagse onderwerpen of onderwerpen die aansluiten bij de leefwereld. Twee van de vier kenmerken van de taakuitvoering richten zich net als de hier beoordeelde Cito toetsen op het begrijpen en interpreteren. De andere twee kenmerken (evalueren van de tekst of informatie mondeling of schriftelijk weergeven bijv. door aantekeningen te maken) komen niet aan de orde, waarschijnlijk omdat er alleen gekozen is om met meerkeuzenvragen te werken. De taken bij de niveaubeschrijving zijn gericht op het luisteren naar instructies, luisteren als lid van een live publiek en luisteren naar radio en tv en naar gesproken tekst op internet. Bij het luisteren naar tv en radio wordt nadrukkelijk aangegeven dat ook via een vooraf gestelde vraag de info uit het verhaal gehaald kan worden. In de toetsen begrijpend luisteren in groep zes vindt dit voor het eerst plaats. In de Cito toetsen begrijpend luisteren wordt alleen naar audiovisuele fragmenten geluisterd (kinderen leven in een beeldcultuur volgens de toetsontwikkelaars) maar niet naar radiofragmenten.

Overall wordt geconstateerd dat wat er getoetst wordt aan begrijpend luisteren en de genoemde inhoudsaspecten op de toetsen begrijpend luisteren dit valide is (zie nog wel aantal gedetailleerde opmerkingen bij vragen van een aantal teksten). Omdat er gekozen is voor meerkeuzenvragen is een beperking waarneembaar van wat er zou kunnen worden getoetst in het domein begrijpend luisteren.

Er wordt opgemerkt dat bij de audio-visuele fragementen er diverse fragmenten zijn waar maar over een beperkt deel vragen worden gesteld. Daarnaast zou het toch ook te overwegen zijn niet alleen te kiezen voor audiovisuele fragmenten, maar ook een fragmenten zonder beeld in te zetten: luisteren pur sang. De fragmenten in de toets die op de uitvoering van liedjes uit Kinderen voor Kinderen gebaseerd zijn, zouden bijvoorbeeld zonder beeld ook bruikbaar zijn.

Merk op dat de beoordeling van dit aspect zich hieronder beperkt tot het statistisch/psychometrisch onderzoek dat is verricht:

De toets Begrijpend Luisteren is niet bedoeld voor voorspellend gebruik. Daarmee is de criteriumvaliditeit niet van toepassing. De (psychometrische) begripsvaliditeit wordt uitgewerkt in unidimensionaliteit, itemkwaliteit, convergente en discriminante validiteit, itembias en in verschillen tussen relevante subgroepen.

De geslaagde kalibratie maakt duidelijk dat het aannemelijk is dat er sprake is van unidimensionaliteit en dat de gekalibreerde itembank één latente trek meet. De gemiddelde moeilijkheidsgraad voldoet met een gemiddelde P van 0,66 voor M4 en 0,74 voor E4 aan het gestelde doel en de moeilijkheidsgraad van de items heeft een goede spreiding. De gemiddelde Rit waarden zijn voor zowel M4 als E4 te kenschetsen als goed.

De constructvaliditeit is uitgewerkt in convergente ('de samenhang tussen de resultaten van het oorspronkelijke onderzoek en de resultaten van een gelijksoortig onderzoek') en discriminante / soortgenoot ('de samenhang tussen de resultaten van het oorspronkelijke onderzoek en de resultaten van een ander onderzoek') validiteit. De convergente correlaties bevestigen de verwachting van een hoge correlatie tussen Begrijpend Luisteren en semantische taalonderdelen en de verwachting van een lage correlatie tussen Begrijpend Luisteren en niet-semantische (technische) taalonderdelen. De soortgenoot validiteit is, rekening houdend met het ontbreken van goed vergelijkingsmateriaal, overeenkomstig de verwachtingen. In het onderzoek naar itembias is geen sprake van DIF (Differential Item Functioning) naar sekse. Conform verwachting scoren meisjes iets hoger dan jongens op de toets Begrijpend Luisteren.

Conclusie:
Aan aspect VL1.1 wordt het oordeel '**voldoende**' toegekend.

Het volg-aspect

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

Bevindingen:
Het algemene (inhoudelijke) uitgangspunt voor de toets Begrijpend Luisteren is dat de (latente) vaardigheid Begrijpend Luisteren kan worden opgevat als een unidimensionaal continuüm en dat elke leerling kan worden voorgesteld als een punt op dit continuüm. Hierbij wordt opgemerkt dat de elementen Begrijpen, Interpreteren en Reflecteren als de drie componenten van de latente vaardigheid Begrijpend Luisteren niet opgevat kunnen worden als te isoleren vaardigheden. Daarmee wordt de vaardigheid Begrijpend Luisteren, als interactie tussen deze drie componenten, beschouwd als één unidimensionele vaardigheid.

In het kalibratieonderzoek is de toets tweemaal afgenomen. Uit het kalibratieonderzoek blijkt dat voor beide afnamemomenten de items passen bij het gehanteerde IRT model en dat het model ook past voor de toets M4/E4 als geheel. Dit betekent dat er sprake is van één unidimensionele vaardigheidsschaal waar items, leerlingen en groei op afgebeeld kunnen worden. In de praktijk wordt de toets binnen een leerjaar echter slechts eenmaal afgenomen.

Afhankelijk van het aantal items dat een leerling goed maakt, wordt er een vaardigheidsscore toegekend. Afhankelijk van het aantal items waaruit de toets bestaat, zal deze vaardigheidsscore meer of minder nauwkeurig kunnen worden geschat. Meisjes scoren in alle gevallen iets hoger dan jongens. Vaardigheidsgroei over de groepen 4, 5 en 6 heen wordt separaat vermeld voor jongens en meisjes. Hierbij wordt opgemerkt dat het per leerjaar steeds om verschillende toetsen gaat, die met behulp van IRT op één en dezelfde vaardigheidsschaal zijn gebracht. Na een relatief grote groei tussen M4 en E4 is er in de leerjaren 5 en 6 sprake van een bescheiden groei. Alleen de groei over de jaren heen wordt in de praktijk echter bij gebruik van het LOVS zichtbaar.

Conclusie:
Aan aspect V1.1 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 de wetenschappelijke verantwoording wordt opgemerkt dat na een relatief grote groei tussen M4 en E4, er in de leerjaren 5 en 6 sprake is van een bescheiden groei. In de tabel op pagina 46 wordt de gemiddelde vaardigheidsscore van M4 tot en met E6, de spreiding per afnamemomenten de toename per afnamemoment vermeld. Toegelicht wordt dat de gemiddelde toename steeds aanmerkelijk kleiner is dan de spreiding in vaardigheid binnen de groep op enig afnamemoment. Dat impliceert dat het meerdere keren vaststellen en in die zin volgen van leerlingen binnen een leerjaar weinig zin heeft. Dit laatste rechtvaardigt ook dat voor deze vaardigheid volstaan kan worden met één afnamemoment per leerjaar.

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 Begrijpend luisteren moet beseft worden dat vaardigheidsgroei zich langzaam in de tijd voltrekt.

Conclusie:
Aan aspect V1.2 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:
In de handleiding zijn registratieformulieren opgenomen voor een leerlingrapport, groepsrapport en een alternatief leerlingrapport (voor leerlingen die op een eigen niveau werken). Deze kunnen door de leraar handmatig of digitaal worden ingevuld. De wijze waarop de registratieformulieren zijn vormgegeven met de uitleg erbij geven inzichtelijk het niveau en de groei weer.

In hoofdstuk 7 van de handleiding ('Communiceren over toetsresultaten') wordt beschreven hoe er met de ouders over de toetsresultaten kan worden gecommuniceerd. Met name wordt daarbij gewezen op het leerlingrapport waarin zowel het niveau van de leerling als de progressie van de leerling numeriek en grafisch gepresenteerd worden. Daarnaast wordt de docent gewezen op misverstanden die zich bij ouders kunnen voordoen bij de interpretatie van de niveau-indelingen. Ook moeten de docenten aan ouders het verschil tussen methode-onafhankelijke en methodegebonden toetsen duidelijk maken en de ouders erop wijzen dat de verschillende toetsen de leerling anders (kunnen) beoordelen.

Conclusie:
Aan aspect VL1.2 wordt 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

- Berkel, S. van, Engelen, R., Hilte, M., Wouda, J. & Zanden, M. van der (2015). *Wetenschappelijke verantwoording Begrijpend luisteren groep 4.* Arnhem: Cito
- Cito (2012). *Begrijpend luisteren groep 4.* Arnhem: Cito.