Cito | Primair en speciaal onderwijs

Wetenschappelijke verantwoording Begrijpend luisteren groep 8

Saskia van Berkel, Maartje Hilte, Frans Kamphuis en Mart van der Zanden

zeker weten

Cito Volgsysteem primair en speciaal onderwijs (LVS)

Begrijpend luisteren

Groep 8

Wetenschappelijke verantwoording

Saskia van Berkel
Maartje Hilte
Frans Kamphuis
Mart van der Zanden

Cito
Arnhem, 2017

Inhoud

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de toets Begrijpend luisteren voor groep 8. De toetsen Begrijpend luisteren maken deel uit van de tweede generatie toetsen van het Cito Volgsysteem primair en speciaal onderwijs (LVS) en zijn primair bestemd voor leerlingen in de groepen 3 t/m 8 in het primair onderwijs. Het betreft voor alle leerjaren papieren toetsen[1].

Deze verantwoording biedt tezamen met de inhoud van het toetspakket Begrijpend luisteren voor groep 8 alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit van het betreffende meetinstrument. Het genoemde materiaal maakt een beoordeling van de toetsen Begrijpend luisteren voor groep 8 mogelijk op de volgende aspecten:
Uitgangspunten van de toetsconstructie;
- De kwaliteit van het toetsmateriaal;
- De kwaliteit van de handleiding;
- Normen;
- Betrouwbaarheid;
- Validiteit.

Het laatstgenoemde aspect betreft alleen begripsvaliditeit en géén criteriumvaliditeit. Omdat de toetsen van het Cito Volgsysteem primair en speciaal onderwijs (LVS) niet bedoeld zijn voor 'voorspellend gebruik' is criteriumvaliditeit niet van toepassing.

Het voorliggende document heeft met name betrekking op de uitgangspunten van de constructie (de hoofdstukken 2 en 3), de normen (hoofdstuk 4), de betrouwbaarheid en meetnauwkeurigheid (hoofdstuk 5) en de validiteit (hoofdstuk 6) van de toets Begrijpend luisteren voor groep 8. De kwaliteit van het toetsmateriaal en de handleiding is te bepalen door kennis te nemen van de inhoud van het toetspakket.

[1] Binnen het Cito Volgsysteem primair en speciaal onderwijs worden geen digitale toetsen Begrijpend luisteren uitgebracht.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

In het onderwijs neemt het toekennen van betekenis aan gesproken taal én het adequaat kunnen reageren op gesproken taal een belangrijke plaats in. Deze vaardigheid wordt in het primair onderwijs aangeduid met de term *begrijpend luisteren* (cf. Verhoeven e.a., 2007; Gijsel & Van Druenen, 2011). De opgaven in de toetsen Begrijpend luisteren voor groep 8 van het Cito Volgsysteem primair en speciaal onderwijs (LVS) zijn een operationalisering van deze vaardigheid.
De toetsen Begrijpend luisteren voor groep 3 tot en met 8 zijn bedoeld om vast te stellen hoe goed leerlingen met begrip kunnen luisteren en hoe hun luistervaardigheid zich ontwikkelt van groep 3 tot en met groep 8 (zie verder paragraaf 2.4.1).

2.2 Doelgroep

De toets Begrijpend luisteren voor groep 8 is primair bestemd voor en genormeerd bij leerlingen in groep 8 van het Nederlandse basisonderwijs. De populatieparameters voor de toets zijn zowel op het midden als op het einde van het schooljaar bepaald. Desgewenst kan de toets ook op een ander moment in het schooljaar worden afgenomen, maar dat maakt het moeilijker om uitspraken te doen over het niveau van de leerling ten opzichte van andere leerlingen in Nederland.
De toets is 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 gemiddelde vaardigheidsscores voor leerlingen uit het reguliere onderwijs. Voor deze leerlingen gelden namelijk dezelfde kerndoelen als voor leerlingen in het basisonderwijs, met dien verstande dat leerlingen in het speciaal (basis)onderwijs meer tijd krijgen om de kerndoelen te bereiken. Deze leerlingen kunnen én moeten dus langs dezelfde meetlat gehouden worden als de ‘reguliere’ leerlingen. De leerlingen in het regulier basisonderwijs op wie de normering gebaseerd is, vormen daarmee ook voor de leerlingen in het speciaal (basis)onderwijs een correcte referentiegroep.
De toets kan ook gebruikt worden voor leerlingen in andere leerjaren die werken op het niveau van groep 8. In de handleiding is toegelicht hoe dit toetsen op maat, met behulp van vaardigheidsscores, in zijn werk gaat. Voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften zijn extra aanwijzingen opgenomen. Voor deze leerlingen zijn daarnaast alternatieve rapportageformulieren ontwikkeld.

Voor leerlingen die nog maar pas in Nederland verblijven, is de toets ongeschikt: leerlingen moeten het Nederlands voldoende beheersen om de opgaven te kunnen maken, voordat de toets Begrijpend luisteren bij hen wordt afgenomen. De toets is ook niet geschikt voor leerlingen met gehoorproblemen.

De toets kan worden afgenomen door de leerkracht of IB’er. We gaan daarbij uit van de professionaliteit van de leerkracht/IB’er. Deze wordt geacht in staat te zijn om aan de hand van de aanwijzingen in de handleiding een gestandaardiseerde en ongestoorde toetsafname te realiseren.

2.3 Gebruiksdoel en functie

De toetsen Begrijpend luisteren in het Cito Volgsysteem primair en speciaal onderwijs (LVS) hebben twee doelen: niveaubepaling en progressiebepaling.

Niveaubepaling
De toetsafnames geven de leerkracht informatie over het niveau van de luistervaardigheid van de leerlingen, individueel en als groep. Iedere behaalde vaardigheidsscore kan daartoe normgericht geïnterpreteerd worden op basis van de vaardigheidsverdeling in een adequate referentiegroep (zie paragraaf 4.2).

In de handleiding zijn twee niveau-indelingen opgenomen, waarmee de leerkracht de scores van een leerling kan vergelijken met de resultaten van een omvangrijke en representatieve steekproef uit de populatie.
De leerkracht kan een keuze maken uit:
- de indeling in de niveaus A tot en met E;
- de indeling in de niveaus I tot en met V.

Bij de indeling in de niveaus A tot en met E is de verdeling over de groepen als volgt:

Niveau	%	Interpretatie
A	**25**	De 25% hoogst scorende leerlingen
B	**25**	De 25% leerlingen die net boven tot ruim boven het landelijk gemiddelde scoren
C	**25**	De 25% leerlingen die net onder tot ruim onder het landelijk gemiddelde scoren
D	**15**	De 15% leerlingen die ruim onder het landelijk gemiddelde scoren
E	**10**	De 10% laagst scorende leerlingen

Bij de indeling in de niveaus I tot en met V wordt uitgegaan van vijf groepen van 20%:

Niveau	%	Interpretatie
I	**20**	De leerlingen die ver boven het gemiddelde scoren
II	**20**	De leerlingen die boven het gemiddelde scoren
III	**20**	De leerlingen die gemiddeld scoren
IV	**20**	De leerlingen die onder het gemiddelde scoren
V	**20**	De leerlingen die ver onder het gemiddelde scoren

In de eerste generatie toetsen uit het leerlingvolgsysteem werd uitsluitend de niveau-indeling A tot en met E gehanteerd. In de praktijk kent deze indeling echter een aantal nadelen.
De indeling is asymmetrisch opgebouwd. De niveaugroepen A, B en C bestrijken elk een kwart van de populatie en het vierde kwartiel is opgesplitst in twee subgroepen: D (15%) en E (10%). Bovendien interpreteert een groot aantal leerkrachten niveau C – het middelste niveau – als gemiddeld. Echter, de indeling A tot en met E toont geen gemiddelde groep leerlingen, maar alleen groepen die boven of onder het gemiddelde scoren.
Daarom is bij de tweede generatie van de toetsen Begrijpend luisteren een indeling geïntroduceerd met de niveaus I tot en met V. Deze indeling is symmetrisch opgebouwd (vijf niveaugroepen van ieder 20%) en heeft als voordeel dat er een ‘werkelijk’ middelste niveau onderscheiden wordt, niveaugroep III. In strikt

statistische zin kan echter ook bij niveaugroep III niet over *het gemiddelde niveau* worden gesproken; het is theoretisch immers mogelijk dat bij een scheve verdeling de gemiddelde ruwe score niet eens in een dergelijke (middelste) groep ligt.

Progressiebepaling
Het volgen van leerlingen in hun groei, ook wel aangeduid als progressiebepaling, is een van de belangrijkste functies van het Cito Volgsysteem primair en speciaal onderwijs (LVS). De toetsen van het LVS geven de leerkracht (en ouders en leerlingen zelf) informatie over de ontwikkeling van de vaardigheden van de leerlingen, individueel en als groep, gedurende (vrijwel) de gehele basisschool-periode. De toetsen geven antwoord op vragen als: is er sprake van vooruitgang, achteruitgang of van stabilisering? Is de vooruitgang – gelet op de gemiddelde vooruitgang in de populatie – volgens verwachting?
Om leerlingen te kunnen volgen wordt de betreffende vaardigheid, in dit geval begrijpend luisteren, opgevat als een unidimensionale vaardigheid, of 'latente trek'. Het gehanteerde meetmodel (zie paragraaf 2.4.2) maakt het mogelijk om de scores van een leerling op verschillende toetsen, op verschillende momenten afgenomen, onderling te vergelijken. De ruwe scores op de toetsen (de ruwe score is het aantal opgaven goed) zijn daartoe te transformeren in scores op één vaardigheidsschaal. Deze unidimensionele vaardigheidsschaal die aan de toetsen Begrijpend luisteren ten grondslag ligt, is ontwikkeld met behulp van het *One Parameter Logistic Model* (Verhelst, 1993; Verhelst & Glas, 1995; Verhelst, Glas & Verstralen, 1995).
Het aantal afnamemomenten per jaar (en daaraan gekoppeld het aantal te construeren verschillende toetsen) wordt bepaald door het tempo waarin een vaardigheid gemiddeld gesproken binnen een leerjaar en over de gehele schoolperiode toeneemt. Meestal is er sprake van twee afnamemomenten per leerjaar ('medio' en 'einde' leerjaar, aangeduid als M en E) en twee – bij het betreffende afname-moment passende – toetsen. Elke toets wordt geconstrueerd op basis van een gekalibreerde itembank, waarbij een toets zo wordt samengesteld dat deze naar inhoud en moeilijkheidsgraad optimaal past bij het afnamemoment waarvoor deze bedoeld is.

Voor de vaardigheid die in deze wetenschappelijke verantwoording aan de orde is, begrijpend luisteren, hebben we in proeftoetsingen vastgesteld dat er in de leerjaren 4 tot en met 8 sprake is van een relatief bescheiden gemiddelde vaardigheidsgroei. Dat betekent dat naar onze mening in deze leerjaren kan worden volstaan met één toetsafname per leerjaar; hetzij op het M-moment, hetzij op het E-moment.
We hebben ons daarom beperkt tot de constructie van één toets die voor beide afname-momenten geschikt is. In groep 8 is gekozen voor normering op alléén het M-moment, omdat scholen in groep 8 geen toetsen meer afnemen aan het eind van het schooljaar.
Dat de keuze voor één afnamemoment per leerjaar correct is geweest, blijkt uit onderstaande gegevens over gemiddelde vaardigheid en vaardigheidsgroei. De gemiddelde toename is steeds aanmerkelijk kleiner dan de spreiding in vaardigheid binnen de groep op enig afnamemoment. Soms is de toename niet veel meer dan een kwart van de standaarddeviatie. Bovendien lijkt de gemiddelde toename over een vol jaar gezien (dat wil zeggen M4-M5, E4-E5 et cetera) steeds kleiner te worden (achtereenvolgens 9,1 - 8,2 - 8,1 - 5,2 – 6,1 - 6) tot E6, en daarna te stabiliseren.

Tabel 2.1 Gemiddelde vaardigheidsgroei voor de afnamemomenten M4 tot en met M8

Afnamemoment	Vaardigheidsscore		
	Gemiddelde	SD	Toename
M4	54,1	8,3	---
E4	59,7	8,6	5,6
M5	63,2	8,9	3,5
E5	65,8	8,6	2,6
M6	70,4	8,6	4,6
E6	73,0	9,2	2,6
M7	76,5	9,3	4,6
E7	79,0	9,2	2,6
M8	86,4	9,8	7,4

Dit impliceert dat het meerdere keren vaststellen en in die zin volgen van leerlingen *binnen* een leerjaar voor begrijpend luisteren in de groepen 4 tot en met 8 weinig zin heeft, mede in relatie tot de nauwkeurigheid van de metingen (waarover zo dadelijk meer). De enige uitzondering hierop is groep 8. Waarom er dan toch opnieuw een flinke vaardigheidsgroei optreedt, daarover valt slechts te speculeren. Mogelijk heeft deze te maken kunnen hebben met de 'mentale sprong voorwaarts' die veel leerlingen tijdens de beginnende puberteit lijken te maken (zie ook paragraaf 6.2.5).

Hoe kunnen we de LVS-toetsen Begrijpend luisteren inzetten om de ontwikkeling van leerlingen te volgen in de tijd? Globaal kunnen we de toetsresultaten van leerlingen (of groepen leerlingen) op twee manieren interpreteren:

a. We kunnen de toetsresultaten van een leerling vergelijken met die van andere leerlingen op hetzelfde meettijdstip (afnamemoment).
b. We kunnen de toetsresultaten van dezelfde leerling vergelijken met diens eigen toetsresultaten op eerdere of latere meettijdstippen (afnamemomenten).

Bij beide vergelijkingen maken we gebruik van het feit dat de toetsresultaten door toepassing van het IRT-model (OPLM) in de vorm van vaardigheidsscores afgebeeld kunnen worden op dezelfde vaardigheidsschaal. Dat geldt voor zowel individuele leerlingen als voor (gemiddelde) groepsresultaten. Dit alles wordt in meer detail uitgelegd in de leerkrachthandleiding (zie het hoofdstuk 'Interpreteren en analyseren op leerling- en groepsniveau').

Ad a.
Bij de eerstgenoemde vergelijking worden de prestaties van een leerling vergeleken met de prestaties van de hele populatie op een gegeven afnamemoment. Hoe doet een leerling het, bijvoorbeeld, ten opzichte van de gemiddelde leerling? Voor dit doel is de populatie, op basis van de data die verzameld zijn in het kader van het normeringsonderzoek (zie hiervoor hoofdstuk 4 van deze wetenschappelijke verantwoording), ingedeeld in vaardigheidsniveaus (I-V, A-E). Vaardigheidsniveau I, bijvoorbeeld, bevat de 20% hoogst scorende leerlingen. Door de vaardigheidsscore van een leerling te vergelijken met deze vaardigheids-niveaus (die zijn afgebakend door percentielpunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk zoals "Mariet heeft op afnamemoment medio leerjaar 5 vaardigheidsniveau IV behaald". Voor de leerkracht (en voor Mariet en haar ouders) bevat deze uitspraak waardevolle informatie.
De leerkracht kan op basis hiervan bijvoorbeeld besluiten om Mariet extra lesstof aan te bieden.

Ad b.
Voor het vergelijken ('volgen') van een leerling op twee verschillende tijdstippen komen twee methodes in aanmerking. Bij de eerste methode worden de **vaardigheidsniveaus** op de twee tijdstippen vergeleken:

“medio groep 7 had Mariet vaardigheidsniveau IV en medio groep 8 was het vaardigheidsniveau V”. Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 74, bijvoorbeeld, op tijdstip M7 en vaardigheidsscore 77 op tijdstip M8. Ook hier geldt, net als bij het vergelijken van prestaties met die van andere leerlingen, dat bij eventuele verdere acties van de leerkracht ook andere aspecten moeten worden betrokken.

Bij alle vergelijkingen die mogelijk zijn, zowel die ad a. als die ad b., dienen uitspraken over leerlingen te worden gerelativeerd. In de handleiding is meer informatie te vinden over de wijze waarop de gebruiker dit kan doen. In paragraaf 3.2 en 4.3 van de leerkrachtmap wordt uitleg gegeven over het score-interval en hoe hiermee om te gaan. In paragraaf 3.2 wordt uitleg gegeven over waarom we ‘score-intervallen’ vermelden in de omzettingstabel ‘Van toetsscore naar vaardigheidsscore en vaardigheidsniveau’ en wat deze score-intervallen inhouden. In paragraaf 4.3 wordt onder stap 1d en stap 2b aan de hand van een (voorbeeld)leerling, Ives, een voorbeeld gegeven van hoe het score-interval geïnterpreteerd kan worden en wat er vervolgens met deze wetenschap gedaan kan worden.

Hieronder gaan we vooral in op het belang van de (on)betrouwbaarheid van de afgenomen toetsen hierbij.Voor elke toets geldt dat de vaardigheidsscore die bij een toetsresultaat van een leerling hoort, behept is met een meetfout. Als we rekening houden met die meetfout dan zou het best zo kunnen zijn dat Mariet vaardigheidsniveau III heeft behaald op het eerste tijdstip en niet vaardigheidsniveau IV. Of dat we moeten concluderen dat het verschil in de prestaties op de twee tijdstippen statistisch niet significant is: Mariet is voor- noch achteruitgegaan. In alle gevallen speelt het betrouwbaarheids-interval (BI) rondom de vaardigheidsscore een belangrijke rol. Dat geldt ook voor de indeling in vaardigheidsniveaus. Hoe het BI doorwerkt in de indeling in vaardigheidsniveaus en de verdere gevolgen daarvan wordt beschreven in hoofdstuk 5. Op deze plaats beperken we ons tot de vraag of we de uitspraak kunnen doen dat een leerling of groep (werkelijk) ‘gegroeid’ is. De eenvoudigste manier is om te kijken of de BI’s voor de twee tijdstippen overlappen. Als deze twee BI’s niet overlappen dan is er sprake van een significant verschil in vaardigheid tussen beide tijdstippen. Overlappen ze wel dan is er geen verschil in vaardigheid. Deze eenvoudige manier van vergelijken kan de leerkracht zelf uitvoeren en wordt ook in de handleiding beschreven. We geven hier een voorbeeld. Bij de afname van Begrijpend luisteren M7 behaalde Wout een vaardigheidsscore van 69 met een 67% betrouwbaarheidsinterval van 64-73. Bij de afname van M8 behaalde Wout een vaardigheidsscore van 79; het bijbehorende betrouwbaarheidsinterval daarbij is 74-84. Aangezien de betrouwbaarheidsintervallen niet overlappen kunnen we zeggen dat Wouts vaardigheid is toegenomen.

2.4 Theoretische inkadering

2.4.1 Theoretische inkadering: inhoudelijk

In deze paragraaf wordt toegelicht wat het concept ‘begrijpend luisteren’ inhoudt. Ook komen de vaardigheden ‘Begrijpen’, ‘Interpreteren’ en ‘Reflecteren’ aan bod, evenals de context waarin luisteren plaatsvindt. Verder leggen we de relatie tussen Begrijpend luisteren en enkele andere taalvaardigheden. We bespreken aan de hand van de kerndoelen Nederlands, de tussendoelen en leerstoflijnen de ontwikkeling van de luistervaardigheid. Ten slotte vergelijken we de uitgangspunten voor de toetsreeks Begrijpend luisteren met de inhoud van het referentiekader Taal, subdomein Luisteren.

2.4.1.1 Het concept begrijpend luisteren

Uit diverse theorieën over en onderzoeken naar de luistervaardigheid (vgl. Bostrom, 1997; Buck 1991; Buck 2001; Damhuis & Litjens, 2003; Krom, Ouborg & Kamphuis, 2001; Levelt, 1989; Rost, 1999; Spearitt, 1999) komt naar voren dat luisteren kan worden opgevat als een actief en constructief proces dat betekenis verleent aan gesproken taal. Luisteren is een proces dat zich afspeelt in het hoofd van de luisteraar: de luisteraar luistert naar gesproken taal, herkent de klanken en identificeert deze als linguïstische eenheden, activeert de betekenis ervan en begrijpt en interpreteert deze, waarbij hij gebruikmaakt van de gegeven informatie en van zijn kennis van de wereld. Tegelijkertijd herinterpreteert de luisteraar voortdurend de

betekenis die hij heeft toegekend in het licht van nieuwe informatie die tijdens het luisteren beschikbaar komt en reflecteert hij op wat er gezegd wordt, bijvoorbeeld door de gegeven informatie te vergelijken met zijn eigen kennis en voorkeuren.
Een luisteraar reconstrueert met andere woorden: hij zet reeksen klanken, waarin de bedoeling van de spreker verpakt is, om in inhouden en hij probeert 'opnieuw' een betekenis samen te stellen.
Zijn reconstructie is geslaagd als de 'nieuw' gereconstrueerde betekenis overeenkomt met de betekenis die de spreker voor ogen had. Luisteren is ook een interactief proces, waarbij de nadruk ligt op het gedrag van de luisteraar: op wat de luisteraar als deelnemer van de samenleving doet of zou moeten doen. Bij luisteren gaat het dus niet alleen om het toekennen van betekenis aan gesproken taal, maar ook om het adequaat kunnen reageren op gesproken taal.
Dit valt in grote lijnen samen met het luisteren dat in het onderwijs – naar analogie van de gangbare tweedeling bij lezen – met de term *begrijpend luisteren*[2] wordt aangeduid.

Karakteristieken van gesproken taal
Bij 'luisteren' gaat het om luisteren naar gesproken taal. Gesproken taal kent een aantal belangrijke karakteristieken (Buck, 2001). Zij bestaat op de eerste plaats uit klanken die de luisteraar moet ontsleutelen en herkennen als betekenisdragende elementen, van kleinere en grotere omvang. De kleinste elementen, de fonemen, worden gecombineerd in woorden, zinnen en teksten. Ze veranderen daardoor vaak enigszins van vorm, bijvoorbeeld in de context van andere klanken of ze verdwijnen of assimileren met andere klanken. Desondanks zijn luisteraars in staat de boodschap van de spreker te ontsleutelen. Verschillende mechanismen vergemakkelijken dit. Zo benadrukt klemtoon wat belangrijk is en geeft intonatie aanwijzingen over de structuur en betekenis van een uiting of reeks uitingen. Daarnaast passen sprekers hun taal aan hun gesprekspartner aan: als er sprake is van veel gedeelde kennis, spreken ze sneller en minder gearticuleerd. Als er minder gedeelde kennis is, spreken ze langzamer en benadrukken ze woorden met een hoge informatieve waarde en krijgen overbodige woorden weinig nadruk. Luisteraars maken ook gebruik van hun kennis van de taal om ontbrekende informatie aan te vullen; alle informatie hoeft niet nadrukkelijk geuit te worden. Kortom, luisteraars moeten net voldoende informatie kunnen oppikken om hun kennis te kunnen activeren, de betekenisconstructie doen ze vervolgens zelf.
Gesproken taal kent op de tweede plaats een aantal eigen, heel specifieke linguïstische verschijnselen (vgl. Tannen 1982; Poelmans, 2003). Spreektaal is een relatief autonoom systeem met verschillende functies. Zij is contextafhankelijk, vluchtig, spontaan, redundant en informeel. De meeste gesproken uitingen zijn een min of meer ruwe eerste versie, ze zijn spontaan, niet gepland, en worden geproduceerd zonder veel tijd voor planning en organisatie. Omdat het werkgeheugen een beperkte capaciteit heeft, bestaat gesproken taal uit kleine ideeënneenheden met een eenvoudige grammaticale structuur, die bijeengehouden worden door nevenschikkende verbanden (en, maar, of). Er zijn aarzelingen en pauzes, opvullers en herhalingen die de spreker extra denktijd geven, er zijn verbeteringen (valse starts, correcties in vocabulaire of zinsbouw) en heroverwegingen. Verder kent spreektaal ook verschijnselen die niet tot de standaardtaal behoren, zoals dialect en alledaagse uitdrukkingen.
Op de derde plaats wordt gesproken taal op nagenoeg hetzelfde moment uitgesproken en beluisterd. Dat vraagt veel van de luisteraar. Gesproken taal is immers vluchtig van aard en direct na het luisteren 'verdwenen'. Ook is er niet altijd gelegenheid om te *her*luisteren. Het is dus noodzakelijk dat het luisterproces efficiënt en in hoge mate automatisch verloopt, zodat de luisteraar de benodigde kennis kan activeren en beschikbaar heeft. Na het luisteren kan hij immers alleen nog een beroep doen op zijn geheugen. Ook al zijn luisteraars over het algemeen goed in staat om de boodschap van de spreker te ontsleutelen, soms gaat er nog wel eens iets mis. Zo kan het voorkomen dat een uiting onvolledig wordt opgeslagen in het geheugen van de luisteraar, bijvoorbeeld door achtergrondlawaai, afleiding of gebrek aan aandacht. Ook 'horen' luisteraars soms verschillende dingen; een effect van hun achtergrondkennis en/of verwachtingen. Of hebben ze andere interesses, behoeften of motieven om te luisteren, waardoor ze verschillende dingen onthouden of dingen verschillend onthouden. Hoewel luisteren een individueel proces

[2]In deze verantwoording hanteren we de term 'luisteren' als het gaat om het luisterproces in de algemene zin van het woord en de term 'begrijpend luisteren' als het gaat over het luisterproces dat plaatsvindt in de schoolse context.

is en interpretaties kunnen variëren, destilleren competente luisteraars wel degelijk dezelfde informatie uit expliciete boodschappen en onthouden zij doorgaans dezelfde gemeenschappelijke kern.

2.4.1.2 Begrijpen, Interpreteren en Reflecteren

Bij luisteren is sprake van interactie tussen drie componenten: de luisteraar met zijn *vaardigheden*, de *tekst* en de *context* (Sijtstra, 2005). Wanneer de luisteraar betekenis toekent aan gesproken taal gebeurt dat altijd in interactie met de tekst. De reactie van de luisteraar wordt bepaald door datgene wat de spreker ter sprake brengt, maar ook door de inbreng van zijn 'eigen' kennis en zijn eerdere (luister)ervaringen. Daarnaast is ook het doel dat de luisteraar voor ogen heeft, bepalend voor zijn reactie.
In het toekennen van betekenis aan gesproken taal spelen zowel tekst- als kennisgestuurde verwerkingsprocessen een belangrijke rol. Bij tekstgestuurde verwerking staat de inhoud van de tekst centraal en verwerkt de luisteraar de informatie die de spreker expliciet ter sprake brengt. Krom e.a. (2011) en Sijtstra (2005) duiden tekstgestuurde verwerking ook wel aan met de vaardigheid *Begrijpen.* Om tot begrip van de tekst te komen, maakt de luisteraar gebruik van de inhoud (de betekenis van woorden, woordgroepen, zinnen, langere tekstpassages en hun onderlinge betekenisrelaties), van expliciete relaties tussen elementen in een uiting of tekst (woord- en zinsvolgorde, verwijzingen en talige structuurmarkeerders) en van de expliciete structuur van een tekst (zie ook Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a).
Kennisgestuurde verwerking gaat verder: om tot begrip van de tekst te komen, zet de luisteraar ook 'eigen' *kennis* in, waaronder zijn kennis van de wereld, zijn kennis over taal en zijn kennis over taalgebruikssituaties. De spreker veronderstelt bepaalde kennis bij de luisteraar bekend en zal die kennis niet altijd expliciteren. Het is aan de luisteraar om deze kennis te activeren en aan te vullen met eigen kennis.
Tussen tekstgestuurde en kennisgestuurde verwerking vindt een continue wisselwerking plaats.
Pas wanneer tekst- en kennisgestuurde verwerking in samenhang en gelijktijdig ingezet worden, is er sprake van werkelijk en diepgaand tekstbegrip. Krom e.a. (2011) en de Expertgroep Doorlopende leerlijnen (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a) spreken in dit verband van *Interpreteren*. De luisteraar vult als het ware de informatie die de spreker geeft verder in en aan met kennis uit andere bronnen. Het onderkennen van en afleiden van impliciete informatie in een tekst, met andere woorden: het maken van inferenties, is een belangrijk aspect van deze vaardigheid.
Luisteraars beschouwen en evalueren ook geregeld teksten en elementen daaruit. Ze nemen dan als het ware afstand van datgene wat ze horen, vormen zich er een mening over en/of toetsen die aan een bepaald standpunt. Dit wordt ook wel aangeduid als de vaardigheid *Reflecteren* of *Evalueren.* Het kenmerkende van deze vaardigheid is de beschouwende en kritische kijk op de tekst. Het gaat niet meer om begrip als zodanig, maar om denken over, reflecteren en abstract redeneren. Dit kan uitmonden in uitspraken over de tekst in evaluerende en waarderende zin (cf. Krom e.a., 2011; Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a).
Echter, men moet zich realiseren dat in werkelijkheid de vaardigheden Begrijpen, Interpreteren en Reflecteren niet zo duidelijk van elkaar te scheiden zijn: ze grijpen op elkaar in, beïnvloeden elkaar en bouwen op elkaar voort. Ze kunnen en mogen dan ook niet opgevat worden als te isoleren vaardigheden van het begrijpend luisteren.

2.4.1.3 Begrijpend luisteren in context

De gesprekken waaraan luisteraars deelnemen, vinden veelal plaats in het dagelijks leven in de vorm van dialogen en polylogen. Deze gesprekken kenmerken zich door tweerichtingsverkeer, waarbij interactie optreedt tussen spreker en luisteraar, waarin spreker en luisteraar van rol kunnen wisselen en waarin zowel auditieve als visuele stimuli in het geding zijn. Het verwerven, verwerken en onthouden van informatie in de interpersoonlijke context is hierbij belangrijk. De functie van dit soort gesprekken ligt vooral in het handhaven van sociale relaties, de inhoud doet er minder toe. Van belang is wel dát er iets gezegd wordt. Er is met andere woorden sprake van interactioneel taalgebruik.
Daarnaast is er transactioneel taalgebruik, met als belangrijkste functie informatieoverdracht. Luisteren naar de radio of naar luisterboeken, luisteren naar de uitleg van een leerkracht of ouder, maar ook televisiekijken zijn daar voorbeelden van. Het gaat hier om informatieoverdracht in de brede zin van het woord, gericht op het begrijpen en interpreteren van de inhoud en op het bepalen van een standpunt of het uitvoeren van een

opdracht. Deze vorm van taalgebruik kan ook gericht zijn op het opdoen van literaire ervaringen, zoals dat gebeurt bij het luisteren naar fictie in de vorm van verhalen en luisterboeken. Het verschil met interactioneel taalgebruik betreft vooral de functie: het gaat de spreker om het overdragen van informatie en de luisteraar om het verwerven van informatie.

Luisteren gaat vaak gepaard met kijken; in de huidige maatschappij speelt beeld een steeds belangrijkere rol. In vrijwel alle gesprekssituaties die zich afspelen in het dagelijks leven en via de media, zijn behalve auditieve stimuli ook visuele stimuli in het geding. Alleen *luisteren,* als geïsoleerde bezigheid, komt nauwelijks nog voor. We leven in een 'beeldcultuur' waarin leerlingen veel sterker visueel zijn ingesteld dan voorheen. Taalmethodes en/of leerkrachten maken steeds meer en steeds vaker gebruik van beeld als instructiemateriaal of van beeld om de leerstof te introduceren of te verduidelijken. Ook in de lessen 'luistervaardigheid' wordt in toenemende mate gebruikgemaakt van beeldmateriaal. Of in de evaluatie van de luistervaardigheid door de aanwezigheid van beeld het construct *luistervaardigheid* wezenlijk verandert, daarover geeft de literatuur vooralsnog geen uitsluitsel (Krom e.a., 2011). Krom merkt hierover op: 'Zolang gesproken teksten aan de basis liggen van luistertoetsen, kan aangenomen worden dat – indien bepaalde voorwaarden vervuld zijn – deze toetsen het construct 'aftappen' (cf. Krom, 2011, p. 24).

2.4.1.4 Begrijpend luisteren en andere vaardigheden

Luistervaardigheid is van belang om zowel in de thuisomgeving als op school en daarbuiten goed te kunnen functioneren: veel van wat kinderen leren, verwerven ze immers door te luisteren. Vooral in het begin van het basisonderwijs, als leerlingen nog niet kunnen lezen, is het een belangrijke manier van informatieoverdracht. Daarnaast vormt begrijpend luisteren de basis voor begrijpend lezen (Gijsel & Van Druenen, 2011).
In de hogere leerjaren neemt ook de noodzaak tot zorgvuldig en kritisch luisteren toe en worden er hogere eisen aan de luistervaardigheid van de leerlingen gesteld, onder meer door de introductie van de zaakvakken. Kinderen moeten in de loop van het basisonderwijs steeds complexere teksten leren begrijpen, waaronder verhalende en informatieve teksten over onderwerpen en situaties waarmee ze nog geen ervaring hebben. Deze teksten komen in de bovenbouw veel voor, onder andere in het kader van wereldoriëntatie (Verhoeven e.a., 2007). De relatie met begrijpend lezen en woordenschat tekent zich dan steeds duidelijker af.

Begrijpend luisteren en begrijpend lezen

Tekstbegrip neemt zowel bij begrijpend luisteren als bij begrijpend lezen een centrale plaats in. Leerlingen die lezen moeten net als leerlingen die luisteren kunnen vaststellen waarover de tekst gaat, voor wie deze bedoeld is en wat de schrijver of spreker met zijn tekst wil bereiken.
Daarnaast zijn er grote overeenkomsten in de verwerkingsprocessen van lezers en luisteraars. Zowel lezers als luisteraars moeten de tekst decoderen, begrijpen en interpreteren. Ook het toepassen van linguïstische kennis en het inzetten van achtergrondkennis is zowel bij begrijpend luisteren als bij begrijpend lezen aan de orde.
Maar de beide processen verschillen ook op wezenlijke punten. Het belangrijkste verschil vloeit voort uit de verschijningsvorm van de tekst: de lezer neemt geschreven tekst tot zich, de luisteraar gesproken tekst. De lezende leerling kan tijdens het lezen teruggrijpen naar de tekst door deze te herlezen, terwijl de luisterende leerling – nadat hij de tekst heeft beluisterd en deze is 'verdwenen' – een beroep moet doen op zijn geheugen.
Een ander verschil betreft het reflecteren op gesproken en geschreven teksten. Omdat tijdens het lezen de tekst beschikbaar blijft, is reflectie op de tekst gemakkelijker dan tijdens het luisteren. Vanwege de vluchtige aard van de tekst moet de luisteraar, veel sterker dan de lezer, de binnenkomende informatie snel en vrijwel automatisch verwerken (cf. Buck, 2001; p. 6). Er is geen mogelijkheid om 'even terug te kijken in de tekst' en zelfs als de spreker (een deel van) de tekst herhaalt, zal deze een tweede keer nooit op precies dezelfde wijze uitgesproken worden als de eerste keer.
Uiteraard spelen de specifieke tekst en de context bij de verwerking van een tekst een cruciale rol.
In situaties waarin de luisteraar geheel is 'overgeleverd' aan de spreker, de zogeheten eenrichtingssituaties, heeft hij geen mogelijkheid tot inbreng. Wanneer zich dan begrips- of interpretatieproblemen voordoen,

moet de leerling tegelijkertijd op meerdere niveaus actief zijn door zowel de problemen op te lossen als de draad van het verhaal niet te verliezen. Dit is een zeer complexe opgave. In interactieve situaties ligt dit anders. De leerling kan dan inbreken in het gesprek en zijn begrip proberen bij te stellen als hij de draad dreigt te verliezen.

Begrijpend luisteren en woordenschat
Woorden vervullen een centrale rol bij het verwerven en toegankelijk maken van kennis: vrijwel alle leerstof is verpakt in woorden, leerkrachten geven woord voor woord uitleg, ze verwoorden verklaringen, brengen gedachteprocessen onder woorden en beschrijven verschijnselen en gebeurtenissen die zich elders in de ruimte en de tijd voordoen. Woorden zijn de bouwstenen van de taal en liggen aan de basis van alledaagse en schoolse kennisoverdracht (vgl. onder meer Van den Nulft en Verhallen, 2002; Verhallen en Verhallen, 1994). Leerlingen die beschikken over een ruime woordenschat nemen gemakkelijker en meer mondelinge (en schriftelijke) informatie tot zich dan leerlingen met een beperktere woordenschat. Omdat ze al veel woorden en betekenissen kennen, kunnen ze nieuwe woorden en woordbetekenissen gemakkelijk inpassen in wat ze al weten en zijn ze tijdens het luisteren in staat om de betekenis van onbekende woorden te achterhalen. Op deze wijze leren ze nieuwe concepten en verbreden ze de betekenisnuances van woorden. Dit staat in schril contrast tot leerlingen met een woordenschatachterstand. Voor deze leerlingen geldt dat de tekst die ze horen vaak zoveel onbekende woorden bevat dat ze de betekenis ervan onvoldoende of in het geheel niet kunnen afleiden. Deze leerlingen begrijpen daardoor veel minder goed wat er wordt gezegd, nemen minder informatie tot zich, leren weinig of zelfs geen nieuwe woorden en de kans om achterop te raken is groot. Vanaf de bovenbouw van het basisonderwijs is een brede, oppervlakkige woordkennis niet meer toereikend en is diepe woordkennis noodzakelijk. Leerlingen moeten dan over een uitgebreid begrippennetwerk beschikken en over woordkennis die snel kan worden ingezet om verbanden en principes te begrijpen en problemen te kunnen oplossen.

2.4.1.5 Begrijpend luisteren in het onderwijs: kerndoelen, tussendoelen en leerstoflijnen
Begrijpend luisteren vormt de basis voor begrijpend lezen (Gijsel & Van Druenen, 2011). De ontwikkeling van begrijpend luisteren kan, net als bij begrijpend lezen, gezien worden als een cyclisch, concentrisch proces: leerlingen doorlopen herhaaldelijk dezelfde ontwikkelings- en leerprocessen, maar op een steeds hoger niveau (Sijtstra, Aarnoutse & Verhoeven, 1999, in: Aarnoutse & Verhoeven, 2003). De verschillende aspecten van begrijpend luisteren, zoals het bepalen van het onderwerp van een tekst of het leggen van verbanden, worden dan ook in alle jaargroepen aan de orde gesteld. In de leerlijnen is er sprake van steeds dezelfde hoofdvaardigheden (Begrijpen, Interpreteren, Evalueren) die de leerlingen in steeds complexere situaties toepassen.

Voor het onderwijs in begrijpend luisteren zijn de kerndoelen van het Nederlands voor het basisonderwijs leidend (Ministerie van OCW, 2006). De vaardigheid Begrijpend luisteren is grotendeels ondergebracht bij ‘Mondeling taalonderwijs’, kerndoel 1, en voor een klein deel bij Taalbeschouwing, kerndoel 12 (zie bijlage 1).

De tussendoelen Mondelinge communicatie (Verhoeven, 2007), die beschouwd kunnen worden als markeringspunten in de mondelinge taalontwikkeling, verwijzen naar de kerndoelen. Ze geven aan wat leerlingen in een bepaalde periode moeten bereiken en zijn opgesteld voor de onder-, midden- en bovenbouw van het primair onderwijs.
In het project TULE, Tussendoelen en Leerlijnen (TULE, 2008), zijn de kerndoelen uitgewerkt in inhouden en activiteiten. Ook de kerndoelen die betrekking hebben op begrijpend luisteren zijn hierin uitgewerkt. TULE beschrijft de tussendoelen voor groep 1/2, groep 3/4, groep 5/6 en groep 7/8.

De tussendoelen en leerstoflijnen voor begrijpend luisteren zijn uitgangspunt geweest bij de opzet en ontwikkeling van de toetsreeks voor groep 4 t/m 8. Ze vormen de basis voor de verschillende opgaventypen en inhoudsaspecten die in de toets Begrijpend luisteren zijn opgenomen (zie paragraaf 3.2.1).

2.4.1.6 Begrijpend luisteren en het referentiekader Taal

In het referentiekader is ‘Luistervaardigheid’, de vaardigheid die in de toetsreeks Begrijpend luisteren getoetst wordt, een subdomein van het domein ‘Mondelinge taalvaardigheid’. Voor het basisonderwijs zijn de referentieniveaus 1F en 2F van belang: 1F is het fundamentele niveau dat elke leerling zou moeten beheersen aan het eind van het basisonderwijs, 2F is het streefniveau voor de leerlingen die meer aankunnen dan 1F. De beschrijving van de referentieniveaus 1F en 2F binnen het subdomein Luisteren luidt:

Referentieniveau 1F
Kan luisteren naar eenvoudige teksten over alledaagse, concrete onderwerpen of over onderwerpen die aansluiten bij de leefwereld van de leerling.

Referentieniveau 2F
Kan luisteren naar teksten over alledaagse onderwerpen, onderwerpen die aansluiten bij de leefwereld van de leerling of die verder van de leerling afstaan.

Als we de uitgangspunten van de toetsen Begrijpend luisteren leggen naast de uitgangspunten bij ‘Luistervaardigheid’ in het referentiekader, zien we dat deze wat betreft de ‘Tekstkenmerken’ nauw overeenkomen. Het gaat dan om de lengte en de opbouw van de luisterteksten.
Van de drie ‘Taken’ die opgenomen zijn in het referentiekader, komen in de toetsen Begrijpend luisteren Taak 1 ‘Luisteren naar instructies’ en Taak 3 ‘Luisteren naar televisie’ voor. Taak 2, ‘Luisteren als lid van een live publiek’ valt (vanzelfsprekend) niet in een toets te realiseren. Onder Taak 3 valt, naast het luisteren naar televisie, ook ‘Luisteren naar radio’: deze aanbiedingsvorm komt in de toetsen Begrijpend luisteren niet voor en dit is een bewuste keuze geweest. Leerlingen in de basisschoolleeftijd luisteren immers niet of nauwelijks meer naar de radio (het luisteren naar muziek daargelaten). Er zijn ook geen speciale kinderprogramma’s voor deze leeftijdsgroep meer op de radio. Hoogstens zal een enkele leerling in groep 8 naar programma’s voor volwassenen luisteren.
Er is in de conceptfase overwogen niet alleen te kiezen voor audiovisuele fragmenten, maar ook fragmenten zonder beeld in te zetten. Het is echter nog niet duidelijk of het ‘luisteren naar audio’ en het ‘luisteren naar beeld met audio’ te verenigen valt binnen één toets, zodanig dat de luistervaardigheid op één vaardigheidsschaal te brengen is. Dat zou eerst onderzocht moeten worden.
In het referentiekader staat verder onder Taak 3: ‘Luisteren naar gesproken tekst op internet’. Gesproken tekst op internet en met name het gebruik van internet op basisscholen stond nog in de kinderschoenen in de conceptvormingsfase van de toetsen Begrijpend luisteren. Gaandeweg het constructieproces hebben we naar bronmateriaal gezocht op internet, maar het resultaat was mager. Hoewel er veel ‘gesproken teksten’ te vinden zijn op internet, is het merendeel inhoudelijk gezien niet bruikbaar voor de toetsen Begrijpend luisteren; ook is de beeld- en/of geluidskwaliteit vaak onvoldoende en is het bronmateriaal (de drager of ‘oorspronkelijke opname’) regelmatig niet meer te achterhalen. Wat wel goed bruikbaar was, bleek vaak voor tv te zijn gemaakt. Daaruit blijkt dus ook dat ‘gesproken tekst op internet’ een diffuus begrip is: is materiaal gemaakt voor tv en vervolgens op internet gezet, een ‘internettekst’ geworden? En omgekeerd: met de komst van zelfstandige ‘tv’-makers’ die hun materiaal direct op internet zetten, is dat materiaal eerder ‘internettekst’ of ‘tv-tekst’? Verder wordt een deel van het materiaal dat gemaakt is voor tv, niet meer uitgezonden op tv maar direct op internet geplaatst door de omroepen (denk aan achtergronden bij tv-programma’s of extra uitzendingen). Ook ‘tv’ en ‘internet’ lopen dus door elkaar heen tegenwoordig.
We hebben teksten gezocht via internet in de constructiefase: deze zijn opgenomen in de toets van groep 8.
Ten slotte komen drie van de vier vaardigheden uit het referentiekader, ‘Kenmerken van de taakuitvoering’, nauw overeen met de uitgangspunten van de toetsen Begrijpend luisteren: ‘Begrijpen’, ‘Interpreteren’ en ‘Samenvatten’. ‘Samenvatten’ is in de toetsen anders uitgewerkt dan in het referentiekader. Daar staat ‘Kan aantekeningen maken. Kan de informatie gestructureerd weergeven’. In de toetsen Begrijpend luisteren wordt ‘samenvatten’ niet gemeten door de leerlingen zelf een samenvatting te laten maken, maar door in de vier antwoordalternatieven vier korte samenvattingen van (een deel van) de tekst te geven en de leerling de ‘juiste’ of ‘beste’ samenvatting te laten kiezen. Deze opgaven zijn ingedeeld bij de vaardigheid ‘Interpreteren’, onder het inhoudsaspect ‘globale inhoud’ (zie paragraaf 3.2.1, tabel 3.2).

De vaardigheid ‘Evalueren’, ‘Kan een oordeel (…) verwoorden’, die ook in het referentiekader genoemd wordt, komt niet terug in de toetsen Begrijpend luisteren, zoals ook verantwoord wordt in paragraaf 3.2.1. De gekozen toetsvorm, een toets met gesloten vragen, maakt dit namelijk onmogelijk.

2.4.2 Theoretische inkadering: psychometrisch

In deze paragraaf gaan we allereerst in op de procedures die we bij de constructie van de toets Begrijpend luisteren hebben gehanteerd; zij komen in paragraaf 2.4.2.1 uitvoerig aan de orde. In deze paragraaf zal ook duidelijk worden dat het gehanteerde IRT-meetmodel in deze procedures een cruciale rol speelt. In paragraaf 2.4.2.2 zal op dit meetmodel worden ingegaan.

2.4.2.1 Opgavenbanken en constructieprocedures

Bij de constructie van opgaven wordt in de regel een veelvoud geproeftoetst van het aantal opgaven dat uiteindelijk in de toets wordt ingezet. Er moet immers rekening worden gehouden met uitval, bijvoorbeeld wegens meer of minder triviale fouten in de constructie of extreme moeilijkheid of gemakkelijkheid. Tegelijkertijd ontstaat er op deze manier een overschot aan kwalitatief goede opgaven, die aan de opgavenbank worden toegevoegd. Uit de kwalitatief goede opgaven worden opgaven geselecteerd voor de uiteindelijke toets. Deze worden vervolgens genormeerd in een normeringsonderzoek (zie paragraaf 4.1). Een belangrijk kenmerk van de opgavenbank is dat ze gekalibreerd is met een IRT-model (voor begrijpend luisteren is dit OPLM; Verhelst en Eggen, 1989; zie verder paragraaf 2.4.2.2), waardoor niet alleen de psychometrische kenmerken (parameters) van de opgaven worden geschat, maar waarbij tevens wordt nagegaan of de opgaven van een onderdeel kunnen worden beschreven met een unidimensionale onderliggende vaardigheid.

Opgavenbank

Voor het samenstellen van toetsen voor het primair onderwijs beschikt Cito over opgavenbanken, die zoals gezegd ten grondslag liggen aan onder meer de toetsen in het Cito Volgsysteem primair en speciaal onderwijs (LVS) en de Entreetoetsen. Een opgavenbank is nadrukkelijk niet eenvoudigweg een verzameling opgaven of items waaruit een toetsdeskundige min of meer naar willekeur een aantal items selecteert om een nieuwe toets samen te stellen. Hieronder wordt beschreven wat de vereisten zijn om van een deugdelijke en psychometrisch goed gefundeerde opgavenbank te kunnen spreken.

- *Unidimensionaal continuüm en latente vaardigheid*

Het algemene uitgangspunt is dat de vaardigheid Begrijpend luisteren kan worden opgevat als een unidimensionaal continuüm (de reële lijn), en dat elke leerling voorgesteld kan worden als een punt op die lijn, met andere woorden: als een getal. Het getal drukt de mate van vaardigheid uit, waarbij een groter getal staat voor een grotere vaardigheid. Het doel van de meetprocedure – het afnemen van de toets – is de plaats van de leerling op dit continuüm zo nauwkeurig mogelijk te bepalen. De uitkomst van de meetprocedure bestaat strikt genomen uit twee grootheden: de eerste grootheid is de schatting van de plaats van de leerling op het vaardigheidscontinuüm. De tweede grootheid geeft aan hoe nauwkeurig die schatting is, en heeft dus de status van een standaardfout, te vergelijken met de standaardmeetfout uit de klassieke testtheorie. De antwoorden van een leerling op de items worden beschouwd als indicatoren van de vaardigheid, hetgeen ruwweg betekent dat men verwacht dat alle items in de bank deze zelfde vaardigheid meten. De vaardigheid zelf wordt als niet-observeerbaar beschouwd, en daarom gewoonlijk omschreven als een latente vaardigheid.

- *‘Moeilijkheid’ in de Item Respons Theorie*

Hoewel items (opgaven) dezelfde vaardigheid meten, kunnen ze toch systematisch van elkaar verschillen. Het belangrijkste verschil tussen de items is hun moeilijkheidsgraad. In de klassieke testtheorie wordt de moeilijkheidsgraad uitgedrukt met een zogenaamde p-waarde, de proportie correcte antwoorden op het item in een welbepaalde populatie van leerlingen. In de Item Respons Theorie (IRT) die voor het construeren van de opgavenbanken wordt gebruikt, hanteert men echter een andere definitie van moeilijkheid: ruwweg gesproken is het de mate van vaardigheid die nodig is om het item goed te kunnen

beantwoorden. Dit verschil in definitie van de moeilijkheidsgraad tussen klassieke testtheorie en IRT is uitermate belangrijk: men kan verwachten dat de p-waarde van een item in groep 7 groter zal zijn dan in groep 6. Dit maakt duidelijk dat 'p-waarde' een relatief begrip is: dat de moeilijkheid van een item in een bepaalde populatie aangeeft. 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. Zo kan men ook de uitspraak begrijpen dat in de IRT vaardigheid en moeilijkheid op eenzelfde schaal liggen.

– *Kansmodel*

De ruwe omschrijving van de moeilijkheidsgraad: de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden (zie vorige alinea), behoeft verdere uitwerking. Deze omschrijving kan worden opgevat als een drempel: heeft een leerling een bepaalde vaardigheid, dan is hij in staat het item juist beantwoorden; heeft hij die vaardigheid niet, dan geeft hij (gegarandeerd) een onjuist antwoord.
Deze interpretatie weerspiegelt een deterministische kijk op het antwoordgedrag van de leerling, die echter in de praktijk geen stand houdt. Er volgt namelijk uit dat een leerling die een moeilijk item correct beantwoordt geen fout kan maken op een gemakkelijk(er) item. Daarom wordt in de IRT een kansmodel gebruikt: hoe groter de vaardigheid, des te groter de kans dat een item juist wordt beantwoord. De moeilijkheidsgraad van een item wordt dan gedefinieerd als de mate van vaardigheid die nodig is om met een kans van precies een half, een juist antwoord te kunnen produceren (zie verder ook paragraaf 2.4.2.2 over het meetmodel).

– *Kalibratie*

In het voorgaande zijn nogal wat veronderstellingen ingevoerd (unidimensionaliteit; alle items zijn indicatoren voor dezelfde vaardigheid; kansmodel) waarvan aangetoond moet worden dat ze deugdelijk zijn. Dit 'aantonen' gebeurt met statistische gereedschappen waarop later nog dieper in wordt gegaan. Maar vóór de items in een toets gebruikt kunnen worden, moet ook geprobeerd worden de waarden van de moeilijkheidsgraden te achterhalen. Dit gebeurt met een statistische schattingsmethode die wordt toegepast op de itemantwoorden die bij een steekproef van leerlingen zijn verzameld. Het hele proces van moeilijkheidsgraden schatten en verifiëren of de modelveronderstellingen houdbaar zijn, wordt kalibratie of ijking genoemd. De steekproef van leerlingen (in de boven al aangeduide proeftoets) die hiervoor wordt gebruikt heet kalibratiesteekproef.

– *Afnamedesigns*

Meestal bevat een opgavenbank meer items dan een doorsnee toets, en is het praktisch ondoenlijk om alle items uit de opgavenbank aan alle leerlingen voor te leggen. Elke leerling in de kalibratiesteekproef krijgt derhalve slechts een (klein) gedeelte van de items uit de opgavenbank voorgelegd. Er is dan sprake van een zogenoemd onvolledig design. Dit gedeeltelijk voorleggen van items moet met de nodige omzichtigheid gebeuren. Voor meer informatie over afnamedesigns die voor de kalibratie kunnen worden gebruikt, verwijzen we de geïnteresseerde lezer naar Eggen (1993).

– *Implicaties van een gekalibreerde opgavenverzameling*

Als de kalibratie met succes uitgevoerd is, is het resultaat een zogenoemde gekalibreerde itembank. In het kalibratieproces worden de items die niet passen bij de verzameling uit de verzameling verwijderd.
De opgavenbank bevat voor elk item niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen, en de statistische zekerheid dat alle items dezelfde vaardigheid aanspreken.

Meer over de kalibratieprocedure en een bespreking van de resultaten daarvan voor de toetsen Begrijpend luisteren is te vinden in hoofdstuk 4 over de normering van de toets.

2.4.2.2 Het gehanteerde meetmodel

In de toetsen Begrijpend luisteren is uitsluitend gebruikgemaakt van een op de itemresponstheorie (IRT) gebaseerd meetmodel zoals dat bij Cito gebruikelijk is, namelijk van het One Parameter Logistic Model (OPLM). Wij zullen dit model hieronder bespreken.

OPLM: het One Parameter Logistic Model

IRT-modellen verschillen in een aantal opzichten nogal sterk van de klassieke testtheorie (Verhelst, 1993; Verhelst & Kleintjes, 1993; Verhelst & Glas, 1995). Bij de klassieke testtheorie staan de toets en de toetsscore centraal. Het theoretisch belangrijkste begrip in deze theorie is de zogenoemde ware score, de gemiddelde score die iemand zou behalen als de toets een oneindig aantal keren onder dezelfde condities zou worden afgenomen. In de IRT staat het te meten begrip of de te meten eigenschap centraal. IRT-modellen hebben belangrijke voordelen boven de klassieke testtheorie. Zo is het bijvoorbeeld mogelijk in de onderzoeksfase van de toetsconstructie te werken met een onvolledig design en kunnen item- en populatieparameters onafhankelijk van elkaar worden geschat (voor een overzicht van de voordelen van IRT-modellen boven de klassieke testtheorie verwijzen we naar Hambleton, Swaminathan en Rogers, 1991).
De IRT beschouwt het antwoord op een item als een indicator voor de mate waarin een eigenschap aanwezig is. Het verband tussen eigenschap en itemantwoord is van probabilistische aard en wordt weergegeven in de zogeheten itemresponsfunctie. Die geeft aan hoe groot de kans is op een correct antwoord als functie van de onderliggende eigenschap of vaardigheid. Formeler: zij X_i de toevalsvariabele die het antwoord op item *i* voorstelt. X_i neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord. Als symbool voor de vaardigheid wordt θ (theta) gekozen. De vaardigheid θ is niet rechtstreeks observeerbaar. Dat zijn alleen de antwoorden op de opgaven. Dit is de reden waarom θ een 'latente' variabele wordt genoemd[3]. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 \mid \theta) \tag{2.1}$$

Een IRT-model is een speciale toepassing van (2.1) waarbij aan de functie $f_i(\theta)$ een meer of minder specifieke functionele vorm wordt toegekend. Een eenvoudig en zeer populair voorbeeld is het zogenaamde Raschmodel (Rasch, 1960) waarin $f_i(\theta)$ gegeven is door

$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \tag{2.2}$$

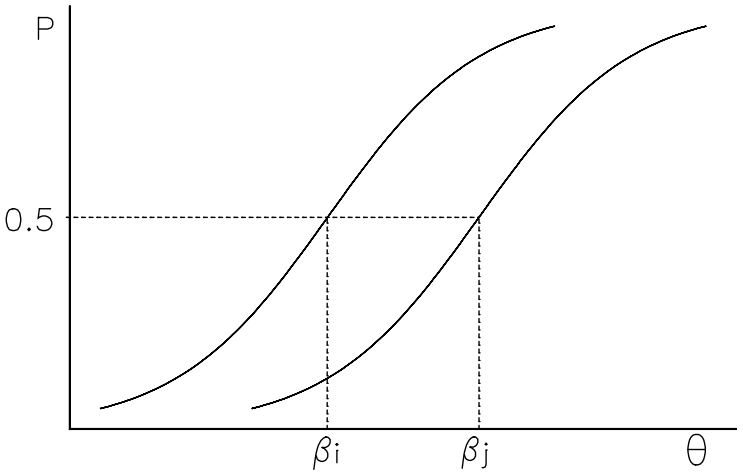
waarin β_i de moeilijkheidsparameter van item *i* is. Dat is een onbekende grootheid die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.1 voor twee items, *i* en *j*, die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende functie is van θ: hoe groter de vaardigheid, des te groter de kans op een juist antwoord. Indien de latente vaardigheid precies gelijk is aan de moeilijkheidsparameter β_i, volgt

$$f_i(\beta_i) = \frac{\exp(\beta_i - \beta_i)}{1 + \exp(\beta_i - \beta_i)} = \frac{1}{1 + 1} = \frac{1}{2} \tag{2.3}$$

Daaruit volgt onmiddellijk een interpretatie voor de parameter β_i: het is de 'hoeveelheid' vaardigheid die nodig is voor de kans van precies een half om het item *i* juist te beantwoorden. Uit de figuur blijkt duidelijk dat voor item *j* een grotere vaardigheid nodig is om diezelfde kans te bereiken, maar dit is hetzelfde als te zeggen dat item *j* moeilijker is dan item *i*. De parameter β_i kan dus terecht omschreven worden als de moeilijkheidsparameter van item *i*. De implicatie van het bovenstaande is dat 'moeilijkheid' en 'vaardigheid' op dezelfde schaal liggen.

[3] Dit maakt duidelijk waarom men de modellen die ressorteren onder de IRT, ook wel aanduidt met 'latente trek'-modellen.

Figuur 2.1 Twee itemresponscurven in het Raschmodel



Formule (2.2) is geen beschrijving van de werkelijkheid, het is een hypothese over de werkelijkheid die getoetst kan worden op haar houdbaarheid. Hoe zo'n toetsing grofweg verloopt, is te verduidelijken aan de hand van figuur 2.1. Daaruit blijkt dat, voor welk vaardigheidsniveau dan ook, de kans om item *j* juist te beantwoorden steeds kleiner is dan de kans op een juist antwoord op item *i*. Hieruit volgt de statistisch te toetsen voorspelling dat de verwachte proportie juiste antwoorden op item *j* kleiner is dan op item *i* in een willekeurige steekproef van personen. Splitst men nu een grote steekproef in bijvoorbeeld twee deel-steekproeven, een 'laaggroep', met de vijftig procent laagste scores, en een 'hooggroep', met de vijftig procent hoogste scores, dan kan men nagaan of de geobserveerde p-waarden van de opgaven in beide deelsteekproeven op dezelfde wijze geordend zijn. Daarvan kan strikt genomen alleen sprake zijn als, in termen van de klassieke testtheorie uitgedrukt, alle opgaven eenzelfde discriminatie-index hebben. Dat echter blijkt lang niet altijd zo te zijn. Ook in ons geval niet. Veel van de items blijken dan ook niet beschreven te kunnen worden met het Raschmodel. Daarom is bij de toets Begrijpend Luisteren gekozen voor een ander IRT-model.

Alvorens dit bij de toets Begrijpend luisteren gebruikte model te introduceren, is eerst een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattings-methode is de 'conditionele grootste aannemelijkheidsmethode' (in het Engels: Conditional Maximum Likelihood, verder aangeduid als CML). Die maakt gebruik van het feit dat in het Raschmodel een afdoende steekproefgrootheid ('sufficient statistic') bestaat voor de latente variabele θ, namelijk de ruwe score of het aantal correct beantwoorde items. Dat betekent grofweg dat, indien de itemparameters bekend zijn, alle informatie die het antwoordpatroon over de vaardigheid bevat, kan worden samengevat in de ruwe score; het doet er dan verder niet meer toe welke opgaven goed en welke fout zijn gemaakt. Hieruit vloeit voort dat de conditionele kans op een juist antwoord op item *i*, gegeven de ruwe score, een functie is die alleen afhankelijk is van de itemparameters en onafhankelijk van de waarde van θ[4]. De CML-schattingsmethode maakt van deze functie gebruik. Deze methode maakt geen enkele veronderstelling over de verdeling van de vaardigheid in de populatie, en is ook onafhankelijk van de wijze waarop de steekproef is getrokken.

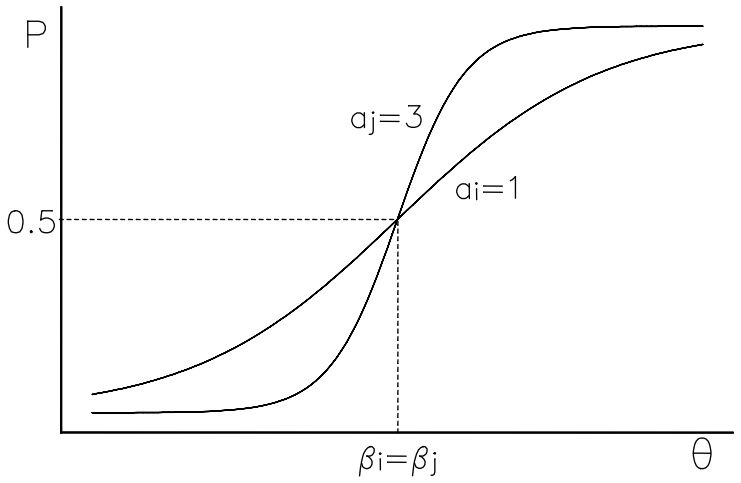
[4] Een gedetailleerde uiteenzetting hierover kan men vinden in Verhelst, 1992.

De CML-schattingsmethode is echter niet bij elk meetmodel toepasbaar. In het zogeheten éénparameter logistisch model (One Parameter Logistic Model, afgekort: OPLM) is CML mogelijk. Dit model is, anders dan het Raschmodel, wel bestand tegen 'omwisseling' van 'proporties juist' in verschillende steekproeven (Glas & Verhelst, 1993; Eggen, 1993; Verhelst & Kleintjes, 1993). De itemresponsfunctie van het OPLM is gegeven door

$$f_i(\theta) = \frac{\exp[\, a_i(\theta - \beta_i)\,]}{1 + \exp[\, a_i(\theta - \beta_i)]}, \qquad (2.4)$$

waarin a_i de zogeheten discriminatie-index van het item is. Door deze indices te beperken tot (positieve) gehele getallen, en door ze a priori als constanten in te voeren, is het mogelijk CML-schattingen van de itemparameters β_i te maken. In figuur 2.2 is de itemresponscurve weergegeven van twee items *i* en *j*, die even moeilijk zijn maar verschillend discrimineren.

Figuur 2.2 Twee itemresponscurven in het OPLM: zelfde moeilijkheid, verschillende discriminatie-index



De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas en Verstralen, 1995). Dit programma voert ook statistische toetsen uit op grond waarvan kan worden bepaald of het model de gegevens adequaat beschrijft. Omdat een aantal van deze toetsen bijzonder gevoelig is voor een verkeerde specificatie van de discriminatie-indices, zijn de uitkomsten van deze toetsen bruikbaar als modificatie-indices: ze geven een aanwijzing in welke richting deze discriminatie-indices moeten worden aangepast om een betere overeenkomst tussen model en gegevens te verkrijgen. Kalibratie van items volgens het OPLM is dan ook een iteratief proces waarin alternerend de modelfit van items wordt onderzocht door middel van statistische toetsing en de waarden van de discriminatie-indices worden aangepast op grond van de resultaten van deze toetsen. Deze aanpassingen geschieden in de praktijk op basis van een en hetzelfde gegevensbestand. Er kan dus kanskapitalisatie optreden. Indien een steekproef een voldoende grootte heeft, is het effect van deze kanskapitalisatie echter gering (Verhelst, Verstralen en Eggen, 1991).

Hoewel het OPLM aanzienlijk flexibeler is dan het Raschmodel, heeft het met dit model toch een nadeel gemeen, waardoor het bij het kalibreren van meerkeuzeopgaven niet zonder meer bruikbaar is. Uit de formules (2.2) en (2.4) volgt dat, indien θ zeer klein is, de kans op een juist antwoord zeer dicht in de buurt van nul komt. Maar de items in het normeringsonderzoek zijn meerkeuze-items, zodat blind gokken een zekere kans op een juist antwoord impliceert. Er bestaan modellen die rekening houden met de raadkans (Lord & Novick, 1968), maar die laten geen CML-schattingsmethode toe. De ongeschiktheid van het Raschmodel of OPLM voor meerkeuzevragen is echter relatief: indien de items in vergelijking met de vaardigheid van de leerling niet al te moeilijk zijn, blijkt dat het effect van het raden op de overeenkomst tussen model en gegevens klein is. Door een verstandige dataverzamelingsprocedure toe te passen en met

name niet te moeilijke opgaven te selecteren in de toets kan het OPLM toch toegepast worden op meerkeuzevragen, waarbij de overeenkomst tussen model en data de uiteindelijke doorslag over die geschiktheid moet geven. Ook in de normering wordt hiermee rekening gehouden.

Voor de schatting van parameters van de populatieverdeling wordt gebruik gemaakt van de ‘marginale grootste aannemelijkheidsmethode’ (in het Engels: Marginal Maximum Likelihood, verder afgekort als MML). Deze schattingsmethode veronderstelt naast (2.2) ook nog dat de vaardigheid θ in de populatie een bepaalde verdeling heeft. De meeste computerprogramma’s die IRT-analyses kunnen uitvoeren, veronderstellen een normale verdeling. Bovendien stelt deze methode de voorwaarde dat de steekproef uit die verdeling die voor de schatting gebruikt wordt een aselecte steekproef is.

3 Beschrijving van de toets

3.1 Opbouw en structuur van de toets

Het toetspakket Begrijpend luisteren voor groep 8 uit het Cito Volsysteem primair en speciaal onderwijs (LVS) bestaat uit een toets M8, die bedoeld is voor afname halverwege het schooljaar (in januari/februari). De toets kent dus één afnamemoment.
Bij speciale leerlingen, die functioneren op een lager niveau, kan de toets van het vorige afnamemoment afgenomen worden. Zo kan een leerling uit groep 8 die moeite heeft met de luisterteksten op het niveau van groep 8, de toets M7/E7 maken. Zie voor uitgebreidere uitleg hierover de handleiding in het toetspakket.

Opbouw
De toets bestaat uit drie delen. Deze dienen bij voorkeur te worden afgenomen op drie verschillende dagdelen, zodat de leerlingen geconcentreerd aan elk deel kunnen werken.
De toets voor groep 8 is als volgt ingedeeld:

Deel 1	10 opgaven	± 30 min
Deel 2	12 opgaven	± 30 min
Deel 3	11 opgaven	± 30 min
Totaal	*33 opgaven*	± 90 min

De leerlingen maken in totaal 33 opgaven bij zes verschillende luisterteksten. Ieder deel bestaat uit twee luisterteksten.

Vorm
De toets voor groep 8 bevat een aantal luisterteksten; dit zijn luisterfragmenten met beeld. De leerlingen kijken en luisteren naar de fragmenten en naar de bijbehorende opgaven. Zowel de luisterfragmenten als de vragen staan op een dvd die klassikaal wordt aangeboden. Daarbij is geen interactie mogelijk; er is met andere woorden sprake van een 'eenrichtingssituatie'. Er is gekozen voor een toets waarin kijk-luisterfragmenten zijn verwerkt, omdat deze aanbiedingsvorm nauw aansluit bij het gegeven dat in de huidige samenleving zowel audio als beeld vaak een rol speelt in het luisterproces (zie paragraaf 2.4.1.3). De vragen in de toets doen echter vooral een beroep op begrip van de *gesproken* tekst: dat is immers de essentie van begrijpend luisteren. Beelden maken de toets niet alleen eigentijdser, maar ook aantrekkelijker voor de leerlingen; er is namelijk gebruikgemaakt van luisterfragmenten uit diverse televisieprogramma's voor de jeugd en uit jeugdfilms. Daardoor is de toets aantrekkelijk (zoals ook blijkt uit de feedback in de evaluatieformulieren) wat de betrokkenheid van de leerlingen vergroot en daardoor het vermogen zich op de toets te kunnen concentreren.

De opgaven in de toets Begrijpend luisteren zijn meerkeuzeopgaven. Hiermee wordt het nakijken en het bepalen van de toetsscore zo eenvoudig en objectief mogelijk gehouden. Elke opgave bevat vier antwoordalternatieven; deze staan in het opgavenboekje en worden voorgelezen op de dvd. Bij de constructie van de opgaven is rekening gehouden met de leesvaardigheid van leerlingen in groep 8.

Afname
De leerkracht neemt de toets klassikaal af aan de hand van een dvd en een afnamekaart met afname-instructies. De afname start met een klassikale instructie en een aantal oefenopgaven. De dvd begint met een kijk-/luistertekst met enkele voorbeeldopgaven, zodat de leerlingen vertrouwd kunnen raken met de verschillende opgaventypen die in de toets voorkomen. De kijk-/luisterfragmenten worden eerst een keer in hun geheel getoond. Daarna worden ze in delen herhaald, steeds gevolgd door één of twee vragen (en een enkele keer drie vragen). De leerlingen zien elke vraag op het beeldscherm en worden daarbij auditief ondersteund, de vraag wordt immers voorgelezen.

De vragen zijn met opzet niet opgenomen in het opgavenboekje. In dat geval zouden de leerlingen de vraag vooraf kunnen lezen en alleen nog maar gericht hoeven luisteren om de vraag te kunnen beantwoorden. Direct na elke vraag volgt er een geluidssignaal (een piep) en moet de leerkracht de dvd-speler op pauze zetten. De leerlingen hebben aansluitend tijd om over hun antwoord na te denken en de vraag te beantwoorden door in hun opgavenboekje de letter voor het gekozen alternatief te omcirkelen.
De antwoordalternatieven staan in het opgavenboekje van de leerlingen en worden voorgelezen, waarbij de leerlingen kunnen meelezen.

Rapportage
De toets Begrijpend luisteren is zowel handmatig als via de computer te scoren en te analyseren. Voor het handmatig nakijken kunnen leerkrachten gebruikmaken van een lijst met goede antwoorden, die in de leerkrachtmap is opgenomen. Indien gewenst kan de leerkracht in het Computerprogramma LOVS de foute antwoorden aanklikken. Het Computerprogramma LOVS geeft dan de juiste score.
Na de toetsafname en de correctie van de leerlingantwoorden kunnen de toetsresultaten verwerkt worden op speciaal ontwikkelde rapportageformulieren. In de hoofdstukken 3 en 4 van de handleiding bij het toetspakket Begrijpend luisteren en in de handleiding bij het Computerprogramma LOVS (zie de module Schoolzelfevaluatie) worden de mogelijkheden besproken om verschillende overzichten te maken, zoals leerlingrapporten, groepsrapporten, dwarsdoorsnedes en trendanalyses. Met behulp van deze overzichten kan de kwaliteit van het gegeven onderwijs zowel op leerling- als op groeps- en schoolniveau geanalyseerd worden.

3.2 Inhoudsverantwoording

Allereerst gaan we in paragraaf 3.2.1 in op de inhoud van de toets Begrijpend luisteren voor groep 8.
We bespreken de tekstsoorten en tekstgenres die in de toets zijn opgenomen. Bij de ontwikkeling van de toets Begrijpend luisteren hebben we de kerndoelen Nederlands voor het primair onderwijs en de tussendoelen en leerstoflijnen van TULE (TULE, 2008) geraadpleegd. Deze hebben we vertaald in een aantal inhoudsaspecten en gerelateerd aan de vaardigheden Begrijpen en Interpreteren. In tweede instantie hebben we onze uitgangspunten gelegd naast het referentiekader Taal (zie paragraaf 2.4.1.5).
In paragraaf 3.2.2 komen de criteria aan bod, zoals we die hebben gehanteerd bij de selectie van opgaven voor het samenstellen van de toets Begrijpend luisteren. De informatie in deze paragraaf vormt een aanvulling op de inhoudsverantwoording die is opgenomen in de handleiding van het toetspakket Begrijpend luisteren groep 8.

3.2.1 De toetsen Begrijpend luisteren: een inhoudsanalyse

Indeling in tekstsoorten en tekstgenres
Het voor de toets gebruikte kijk-/luistermateriaal is afgestemd op leerlingen in groepen 8 van het basisonderwijs (vgl. TULE, 2008). De geselecteerde teksten zijn relatief kort, al zijn de langste teksten in de toets qua luisterduur langer dan in de lagere groepen. De luisterduur per tekst bedraagt maximaal zeven minuten.
De teksten sluiten aan bij de leefwereld van de doelgroep en hebben een aan de leeftijdsgroep aangepaste informatiedichtheid; er wordt niet te veel informatie tegelijkertijd aangeboden. Ze zijn eenvoudig en/of helder van structuur en hebben een duidelijke opbouw. Langere zinnen en samengestelde zinnen komen nu vaker voor in vergelijking met eerdere leerjaren.
In de toets zijn twee teksttypen vertegenwoordigd, fictie en non-fictie, en een aantal tekstgenres.
Elke combinatie van teksttype en tekstgenre heeft specifieke kenmerken wat betreft opbouw, stijl, register, doel, publiek, taalgebruik, conventies, mate van formaliteit en de manier van presenteren.
We hebben de volgende tekstgenres onderscheiden:
- lied (fictie);
- film/drama/verhaal (fictie);
- nieuwsbericht (non-fictie);

- interview/gesprek (non-fictie);
- documentaire/verslag (non-fictie);
- betoog, waaronder ook reclame (non-fictie);
- instructie/uitleg (non-fictie).

Het bleek onmogelijk alle tekstgenres in de toets voor groep 8 op te nemen, omdat de toetsafname in dat geval te veel tijd in beslag zou nemen. We hebben wel geprobeerd om een zo groot mogelijke variatie aan tekstgenres in de toets op te nemen. Sommige teksten zijn ondergebracht bij twee genres omdat deze teksten kenmerken van beide genres hebben. (In de praktijk behoren teksten niet uitsluitend tot één genre, maar bestaan ze uit een mengvorm van deze genres.) Zo zitten er in 'Blind vertrouwen op de geleidehond','Ecologisch netwerk' en 'Conditie' verschillende gedeelten waarin een gesprek voorkomt. Daarom hebben we deze teksten op twee plaatsen vermeld. De volgende genres zijn geselecteerd:

- lied: tekst 'De Neus' (Klokhuis-tv/internetsite Klokhuis);
- documentaire/verslag: de teksten 'Ecologisch netwerk (Klokhuis-tv) en 'Blind vertrouwen op de geleidehond' (internet);
- instructie/uitleg: de tekst 'Conditie'(Beeldbank Schooltv via internet);
- betoog/reclame: de teksten 'Katja's stichting' (Beeldbank Schooltv via internet) en 'Tropenmuseum' (internet);
- interview/gesprek: (enkele gedeeltes van) tekst 'Ecologisch netwerk' (Klokhuis-tv), tekst 'Blind vertrouwen op de geleidehond' (internet) en tekst 'Conditie' (Beeldbank Schooltv via internet).

Indeling in opgaventypen naar vaardigheden en inhoudsaspecten
Zoals uiteengezet in paragraaf 2.4.1.2 zetten luisteraars tijdens het luisteren naar gesproken taal een aantal specifieke vaardigheden in. Bij het toekennen van betekenis aan hetgeen ze horen doen ze een beroep op de vaardigheden Begrijpen, Interpreteren en Reflecteren.
Het toetsen van vaardigheden in het reflecteren is – zeker in de voor LVS gekozen toetsvorm, namelijk gesloten vragen – lastig te realiseren. In open vragen zou dit kunnen, maar dan blijft de beoordeling lastig: een mening is immers een mening, er is geen 'goed' of 'fout'. Er kan hoogstens beoordeeld worden of er sprake is van een mening.
We hebben er daarom voor gekozen om in de toets Begrijpend luisteren voor groep 8 alleen opgaven op te nemen die een beroep doen op de vaardigheden Begrijpen en Interpreteren. Het zijn ook met name deze vaardigheden waarop (jeugdige) luisteraars een beroep doen tijdens het luisteren naar gesproken taal.
De verschillende inhouden die in het luisteronderwijs aan bod komen (vgl. de kerndoelen Nederlands voor het primair onderwijs en de tussendoelen en leerlijnen van TULE (TULE, 2008)) hebben we vertaald in een aantal inhoudsaspecten en gerelateerd aan de beide vaardigheden. Op die manier zijn we gekomen tot een aantal verschillende opgaventypen. Elk opgaventype representeert een of meerdere inhoudsaspecten.

Het overzicht op de volgende pagina maakt voor elk van de beide vaardigheden inzichtelijk welke inhoudsaspecten aan welke vaardigheid gerelateerd zijn. Voor voorbeeldopgaven bij elk opgaventype verwijzen we naar hoofdstuk 6 van de handleiding bij de toets.

Begrijpen

opgaventype ‘expliciete betekenistoekenning’
Opgaven die vragen naar de betekenis van een woord dat of woordgroep die expliciet door de spreker vermeld wordt.

opgaventype ‘specifieke inhoudselementen’
Opgaven die vragen naar specifieke inhoudselementen die expliciet in de tekst aan de orde gesteld worden. Dit zijn bijvoorbeeld (hoofd)personen, feiten en meningen, voorwerpen, aantallen, een plaats van handeling of tijdsperioden.

opgaventype ‘eenvoudige expliciete verbanden’
Opgaven die vragen naar eenvoudige expliciete verbanden op basis van inhoudelijke en structurele elementen. Voorbeelden daarvan zijn vergelijkingen, tegenstellingen, generalisaties en voorbeelden, vraag en antwoord. Ook verwijzingen en verbanden tussen kleine stukjes informatie die expliciet verwoord worden – terwijl het verband zelf niet geëxpliciteerd is – en expliciete verbanden die de spreker legt tussen gebeurtenissen, personen of plaatsen zijn hier voorbeelden van.

opgaventype ‘complexe expliciete verbanden’
Opgaven die vragen naar complexe expliciete verbanden op basis van inhoudelijke en structurele elementen over grotere tekstdelen heen. Dit zijn bijvoorbeeld: reden en verklaring, oorzaak en gevolg, middel en doel, deel-/geheelrelaties, conclusie en argumenten, generalisaties en voorbeelden of hoofd- en bijzaken. Maar ook opgaven die vragen naar de chronologie van gebeurtenissen of naar opeenvolgende stappen vallen hieronder.

Interpreteren

opgaventype ‘impliciete betekenistoekenning’
Opgaven die vragen naar het afleiden van de betekenis van een woord of woordgroep.

opgaventype ‘inzet van voorkennis’
Opgaven die vragen naar het afleiden van informatie uit de tekst, waarbij de leerling zijn voorkennis moet inzetten, naast de informatie die de spreker geeft. Het gaat dan om opgaven waarbij ontbrekende informatie moet worden aangevuld, waarbij moet worden geanticipeerd of waarbij naar de bedoeling, gevoelens of mening van de spreker gevraagd wordt. Maar ook opgaven die vragen naar de functionele betekenis van tekstdelen vallen hieronder.

opgaventype ‘globale inhoud’
Opgaven die vragen naar de globale inhoud van de tekst of een tekstdeel waarbij expliciete en/of impliciete inhoudelijke en/of structurele elementen verspreid over de tekst of over grotere tekstdelen, moeten worden verbonden. Voorbeelden hiervan zijn opgaven over onderwerp, thema, hoofdlijnen, hoofdgedachte, hoofdpersoon en doel en publiek van de tekst. Ook opgaven waarbij de leerlingen informatie in de tekst moeten vergelijken en/of doorzien of waarbij ze de inhoud van de tekst of een tekstdeel moeten samenvatten of de opbouw van een tekst moeten doorzien, zijn hier voorbeelden van.

opgaventype ‘manier van spreken’
Opgaven die vragen naar de manier van spreken, bijvoorbeeld naar: klemtoon, intonatie, volume, tempo, toon, accent, register, sociale en culturele conventies en waarbij een verband gelegd moet worden tussen tekstuele informatie en kennis van het taalsysteem.

Verdeling van de opgaven over de toetsen

Bij het samenstellen van de toetsen zijn we ervan uitgegaan dat de vaardigheden Begrijpen en Interpreteren beide een belangrijke rol in het luisterproces vervullen. Omdat in de bovenbouw van het primair onderwijs steeds meer van de leerlingen gevraagd wordt (informatieverwerking bij de zaakvakken bijvoorbeeld) en er in toenemende mate een beroep gedaan wordt op de vaardigheid 'Interpreteren', hebben we ernaar gestreefd het percentage opgaven Interpreteren in de toetsen in de loop van de leerjaren op te hogen. De meest wenselijke verdeling in de bovenbouw is ongeveer een derde Begrijpen en twee derde Interpreteren voor de toets als geheel, zodat Interpreteren in ieder geval meer aan bod komt in de bovenbouw.

Tabel 3.1 Aantal opgaven Begrijpen en Interpreteren in de toets Begrijpend luisteren voor groep 8

Toets	Aantal opgaven Begrijpen	Aantal opgaven Interpreteren	Totaal aantal opgaven
M8	11 (33%)	22 (67%)	33

Tabel 3.1 laat zien dat we hierin redelijk goed zijn geslaagd: 33% van alle opgaven doet een beroep op de vaardigheid Begrijpen en 67% op de vaardigheid Interpreteren. Daarbij moet in aanmerking worden genomen dat de opgaven gecategoriseerd zijn naar de mate waarin het accent meer op het 'Begrijpen' of het 'Interpreteren' ligt; dat betekent dat opgaven dus een beroep kunnen doen op beide vaardigheden. Door de onderscheiden opgaventypen en onderliggende inhoudsaspecten te verdelen over de vaardigheden Begrijpen en Interpreteren, hebben we ons er tijdens de constructiefase van verzekerd dat de luistervaardigheid in al haar facetten en van alle kanten belicht werd. Het bleek in deze fase niet mogelijk om bij elke tekst alle beschikbare opgaventypen en inhoudsaspecten in te zetten, omdat niet alle tekstgenres zich daar even goed voor lenen. Bovendien is in de fase van proeftoetsing een aantal opgaven uitgevallen. Desondanks sluit de verdeling in inhoudsaspecten redelijk goed aan op de verdeling die we beoogden. Alle inhoudsaspecten op een na ('Opgaven die vragen naar specifieke inhoudselementen die expliciet in de tekst aan de orde gesteld worden.') zijn in voldoende mate in de toetsen vertegenwoordigd. Tabel 3.2 geeft de uiteindelijke verdeling weer van de opgaven over de onderscheiden vaardigheden en inhoudsaspecten. Ook is aangegeven welke opgaven tot welk type behoren. De opgaven aangegeven met een asterisk betreffen de twijfelgevallen. Er zijn zowel opgaven opgenomen die een beroep doen op de vaardigheid Begrijpen als op de vaardigheid Interpreteren, en wel in de beoogde verhouding.

Het rubriceren van de opgaven is onafhankelijk en door meerdere toetsdeskundigen gebeurd, waarna de indelingen bekeken zijn en er bediscussieerd is bij welk opgaventype een opgave uiteindelijk moest worden ingedeeld. Echte twijfelgevallen hebben we uiteindelijk toegewezen aan een bepaalde categorie. In tabel 3.2 zijn de opgaven waarover twijfel was, aangegeven met een asterisk. Deze opgaven maken duidelijk dat het soms lastig is te scheiden waar het 'begrijpen' ophoudt en het 'interpreteren' begint. Wij willen met de verdeling van de opgaven over de beide vaardigheden dan ook geenszins suggereren dat dit 'apart te onderscheiden of te oefenen vaardigheden' zijn.

In tabel 3.2 is te zien dat de derde categorie van Begrijpen en de tweede en derde categorie van Interpreteren in verhouding meer gevuld zijn dan de eerste en laatste categorieën: dit is logisch omdat dit brede categorieën zijn waaronder meerdere opgaven vallen en waarvoor ook meer opgaven geconstrueerd zijn. Onder Interpreteren/'Opgaven die vragen naar de manier van spreken' vallen bijvoorbeeld opgaven die vragen naar klemtoon of intonatie en dat betreft een minder omvangrijke categorie dan bijvoorbeeld Interpreteren/'Opgaven die vragen naar het afleiden van informatie uit de tekst', waarbij de leerling zijn voorkennis moet inzetten, naast de informatie die de spreker geeft.

De toets voor groep 8 bevat geen opgaven die vallen in de categorie 'Begrijpen/Opgaven die vragen naar specifieke inhoudselementen' (zoals opgaven die vragen naar persoon, voorwerp, aantal, plaats van handeling, tijd.) omdat de opgaven uit deze categorie veel te eenvoudig bleken te zijn voor de leerlingen in groep 8. De exacte verdeling van inhoudsaspecten over de vaardigheden is niet zo belangrijk. De verdeling werd voornamelijk bepaald door de aard van de teksten en het soort opgaven dat na proeftoetsing op psychometrische gronden konden worden behouden. Hierbij moet in gedachten gehouden worden dat de

vaardigheden Begrijpen en Interpreteren én de diverse inhoudsaspecten in werkelijkheid niet zo heel duidelijk van elkaar te scheiden zijn (vgl. paragraaf 2.4.1.2). We kunnen ze dan ook niet opvatten als te isoleren vaardigheden en aspecten van het begrijpend luisteren. Het feit dat de opgaven op één vaardigheidsschaal liggen, illustreert dit ook.

Tabel 3.2 Verdeling van de opgaven naar vaardigheid en inhoudsaspecten in de toets M8

Vaardigheid en inhoudsaspecten	**Aantal opgaven (opgavennummers[1])**
Begrijpen	
Opgaven die vragen naar de betekenis van een woord of woordgroep die expliciet door de spreker vermeld wordt.	3 (opg. 4, 20, 26)
Opgaven die vragen naar specifieke inhoudselementen die expliciet in de tekst aan de orde gesteld worden.	0
Opgaven die vragen naar eenvoudige expliciete verbanden op basis van inhoudelijke en structurele elementen.	7 (opg. 13, 14, 15, 17, 19, 24, 29)
Opgaven die vragen naar complexe expliciete verbanden op basis van inhoudelijke en structurele elementen over grotere tekstdelen heen.	1 (opg. 12)
Totaal	*11 (33%)*
Interpreteren	
Opgaven die vragen naar het afleiden van de betekenis van een woord of woordgroep.	2 (opg. 3, 10)
Opgaven die vragen naar het afleiden van informatie uit de tekst, waarbij de leerling zijn voorkennis moet inzetten, naast de informatie die de spreker geeft.	9 (opg.7, 8, 21, 25, 27, 28, 31*, 32, 33)
Opgaven die vragen naar de globale inhoud van de tekst waarbij expliciete en/of impliciete inhoudelijke en/of structurele elementen verspreid over de tekst of over grotere tekstdelen, moeten worden verbonden.	9 (opg. 1, 2, 5, 6*, 11, 16, 18, 23, 30)
Opgaven die vragen naar de manier van spreken.	2 (opg. 9, 22)
Totaal	*22 (67%)*

[1] De opgaven van de toets zijn in deze tabel doorgenummerd.

Opgave 1 van deel 2 is hier opgavenummer 11, opgave 2 van deel 2 is opgavenummer 12 etc.

Opgave 1 van deel 3 is hier opgavenummer 23, opgave 2 van deel 3 is opgavenummer 24 etc.

* Opgaven die ook bij een ander itemtype ingedeeld hadden kunnen worden.

3.2.2 Selectie van de opgaven

Alle opgaven die in de toets Begrijpend luisteren zijn opgenomen zijn speciaal voor de toets geconstrueerd door een constructieteam, bestaande uit (oud-)leerkrachten uit het basisonderwijs en een PABO-docent, aan de hand van aanwijzingen van toetsdeskundigen van Cito wat betreft de selectie van de teksten en de constructie van de opgaven. Allereerst zijn in een landelijk proefonderzoek opgaven voorgelegd aan basisschoolleerlingen in groep 8, waarbij het streven was dat elke opgave door minimaal 300 leerlingen gemaakt werd. Het primaire doel van dergelijke proefafnames is het verkrijgen van informatie over de moeilijkheid van de afzonderlijke opgaven. Ook kunnen opgaven met een laag discriminerend vermogen geïdentificeerd en verwijderd worden. Dit zijn opgaven die geen of onvoldoende onderscheid maken tussen vaardigere en minder vaardige leerlingen. Daarnaast biedt een proefafname de mogelijkheid om aan de deelnemende leerkrachten te vragen of ze inhoudelijke of andersoortige bezwaren hebben tegen de aangeboden kijk-/luisterfragmenten of opgaven.
De opgaven die zowel psychometrisch als inhoudelijk geschikt bleken, zijn vervolgens opgenomen in de toets ten behoeve van de normeringsonderzoeken. In principe kwamen alle opgaven met een acceptabele moeilijkheid en een acceptabel discriminerend vermogen hiervoor in aanmerking. Echter, naast psychometrische criteria waren ook inhoudelijke criteria bij de opgavenselectie van belang. Zo wilden we de opgaven zo evenwichtig mogelijk verdelen over de vaardigheden Begrijpen en Interpreteren en over de diverse inhoudsaspecten, maar ook over de diverse tekstgenres. Van alle opgaven die zijn meegegaan in het normeringsonderzoek zijn de gekalibreerde p-waarde, de r_{it}-waarde en de r_{ir}-waarde bepaald. Uiteindelijk zijn er 33 opgaven in de toets M8 opgenomen.
De teksten en opgaven in de toets Begrijpend luisteren zijn gescreend door deskundigen uit het speciaal (basis)onderwijs. Hierbij is erop gelet dat de opgaven geschikt zijn voor een zo groot mogelijke groep leerlingen, ook leerlingen met extra onderwijsbehoeften. Deze screening bleek echter lastig, omdat veel van deze leerlingen het niveau van groep 8 niet halen. Ook is het aandeel Interpreteren en figuurlijk taalgebruik in de toets voor groep 8 in verhouding groot, en zijn de onderwerpen redelijk abstract en de teksten van een behoorlijk hoog niveau.
De deskundigen vonden de toets geschikt voor leerlingen in het speciaal (basis)onderwijs, en hadden slechts bij enkele opgaven bezwaren. Bij de vierde opgave van de tekst ‘De Neus’ vonden ze het verband dat de leerlingen moeten leggen te complex en bij de vijfde opgave bij diezelfde tekst vonden ze het figuurlijk taalgebruik te moeilijk voor ‘hun’ leerlingen. Ook de laatste opgave bij de tekst ‘Katja’s stichting’, een opgave die vraagt naar ‘de manier van spreken’, vonden de deskundigen erg moeilijk voor bepaalde leerlingen, al gaven ze wel aan ‘dat dit wel iets is wat de leerlingen zouden moeten kunnen aan het eind van het basisonderwijs. Ook opgaven die figuurlijk taalgebruik bevragen, konden volgens de deskundigen niet weggelaten worden in een toets Begrijpend luisteren.

3.3 Statistische beschrijving

3.3.1 Itemkenmerken: moeilijkheidsgraad en interne consistentie

Wat de moeilijkheid van de opgaven betreft: voor de opgavenselectie geldt het uitgangspunt dat de p-waarden bij voorkeur tussen 0,40 en 0,90 moeten liggen en dat de opgaven van Begrijpend luisteren gemiddeld een p-waarde tussen de 0,65 en 0,75 hebben. In tabel 3.3 rapporteren we de geschatte range van p-waarden en de geschatte gemiddelde p-waarde van de opgaven voor het meetmoment M8 van de toets Begrijpend luisteren voor groep 8. Daarnaast zijn ook gegevens opgenomen over de R_{it}-waarden van de opgaven, waarbij de toetsscore over het betreffende onderdeel het uitgangspunt was voor de berekening van de coëfficiënt. R_{ir}-waarden zijn wellicht te prefereren omdat zij een realistischer beeld geven van de correlatie met de schaalscore, maar helaas zijn ons geen normgegevens bekend voor R_{ir}. Voor R_{it}-waarden kent het COTAN-beoordelingssysteem (COmmissie TestAangelegenheden Nederland van het Nederlands Instituut van Psychologen, zie Evers et al., 2010) wél kwaliteitscriteria.

Voor de de toets M8 blijken de p-waarden goed in de buurt te komen van de gekozen uitgangspunten. Het gemiddelde ligt op 0,73. Voor het minimum geldt dat geen p-waarde onder de 0,40 uitkomt. Alle p-waarden voor meetmoment M8 liggen onder de 0,90. De gemiddelde R_{it}-waarden zijn voor de toets van groep 8 te kenschetsen als 'goed' (gemiddelde R_{it} > 0,30). Van de in totaal 33 opgaven hebben acht opgaven op meetmoment M8 een Rit-waarde tussen de .20 en .30 (zie bijlage 2 'Items en waarden toets M8', waarin ook de Rir-waarden opgenomen zijn).
Alle andere Rit-waarden liggen boven de 0,30 en zijn te kenschetsen als 'goed'.

Tabel 3.3 Range en gemiddelde van p- en R_{it}-waarden voor de toets M8

	P-waarden		R_{it}-waarden		N items
	Range	**Gem.**	**Range**	**Gem.**	
M8	0,43 - 0,90	0,73	0,23 - 0,47	0,33	33

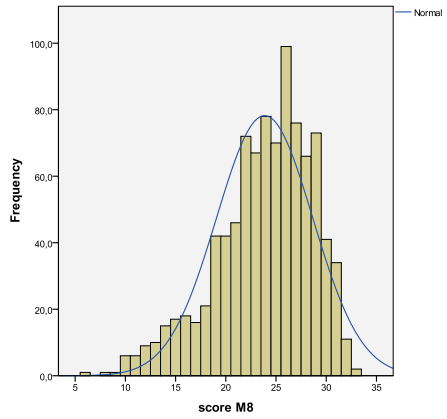
3.3.2 Verdeling van de ruwe scores

In tabel 3.4 zijn de verdelingskarakteristieken gegeven van de ruwe scores op toetsmoment M8.
De gemiddelden komen uiteraard overeen met wat men bij een gegeven aantal items mag verwachten bij de gekozen (gemiddelde) moeilijkheidsgraad. De gemiddelde moeilijkheidsgraad ligt op 0,73, daarom is de verdeling linksscheef (de negatieve waarde in de kolom 'skewness'). De verdeling is ééntoppig. Voor een grafische weergave zie het histogram van de scores in figuur 3.1.

Tabel 3.4 Verdelingskenmerken van het toetsmoment M8

Meetmoment	Aantal opgaven	M	SD	Skewness	Kurtosis
M8	33	23,8	4,79	-0,733	0,275

Figuur 3.1 Histogram van de toetsscores op het afnamemoment M8 voor de toets M8



4 Kalibratie en normering

4.1 Opzet voor de normeringsonderzoeken van het LVS: het macrodesign

Het opzetten van een leerlingvolgsysteem in het basisonderwijs is een complexe onderneming, en het verzamelen van de gegevens om het systeem te ijken en normeren moet met de nodige zorg gebeuren. Immers, het is niet voldoende om voor elke halfjaargroep (M3, E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8) over normen te beschikken, er moet ook voor gezorgd worden dat de prestaties over de jaren heen met elkaar vergelijkbaar zijn. Hiertoe dienen de prestaties van leerlingen over alle leerjaren heen te worden afgebeeld op een gemeenschappelijke vaardigheidsschaal. Om zo'n gemeenschappelijke schaal te realiseren kunnen we niet volstaan met het ontwikkelen van afzonderlijke toetsen voor de meetmomenten en elke toets afzonderlijk ijken en normeren. Prestaties van bijvoorbeeld de populatie M7 moeten vergelijkbaar zijn met die van andere afnamemomenten, bijvoorbeeld E6 en E7. Met andere woorden, het dataverzamelingsdesign dient verbonden te zijn. Hiertoe dient een longitudinale opzet gebruikt te worden.

De verbondenheid van het design
Het idee van een gemeenschappelijke schaal impliceert strikt genomen dat men iemands vaardigheid zou kunnen schatten aan de hand van een willekeurig samengestelde toets. Het spreekt echter vanzelf dat het een zinloze onderneming is een toets die geconstrueerd is voor leerlingen in groep 8 voor te leggen aan leerlingen van groep 3, omdat zo'n toets ongetwijfeld opgaven zal bevatten die een beroep doen op kennis van leerstof die in groep 3 niet is onderwezen. Dit betekent dat we door de algemene kenmerken van het curriculum tamelijk beperkt zijn in het voorleggen van itemmateriaal aan leerlingen voor wie het niet specifiek is geconstrueerd. Daarom is er besloten dat het overlapmateriaal dat aan een bepaalde (half-)jaargroep kan worden voorgelegd alleen itemmateriaal mag bevatten dat specifiek voor die halfjaargroep is geconstrueerd en voor de twee belendende halfjaargroepen. Voor M8 betekent dit dat de leerlingen in het kalibratie- en normeringsonderzoek items voorgelegd krijgen die specifiek voor M8 zijn geconstrueerd, en (een minderheid aan) items die geconstrueerd zijn voor E7. Het macrodesign is weergegeven in onderstaande figuur.

Figuur 4.1 Macrodesign LVS Begrijpend luisteren

	E2	M3	E3	M4	E4	M5	E5	M6	E6	M7	E7	M8
jan 2011 M3	ank23	M3	ank33									
mei 2011 E3		ank33	E3	ank34								
jan 2012 M4			ank34	M4	ank44							
mei 2012 E4				ank44	E4	ank45						
jan 2013 M5					ank45	M5	ank55					
mei 2013 E5						ank55	E5	ank56				
jan 2014 M6							ank56	M6	ank66			
mei 2014 E6								ank66	E6	ank67		
jan 2015 M7									ank67	M7	ank77	
mei2015 E7										ank77	E7	ank78
jan 2016 M8											ank78	M8

De items die voor de overlap of verankering zorgen, duiden we in het macrodesign aan met ank, gevolgd door 2 cijfers. Zo duidt ank67 de groep items aan die enerzijds bestaat uit items geconstrueerd voor E6 en anderzijds uit items geconstrueerd voor M7. Die items zijn dus zowel eind groep 6 als medio groep 7 afgenomen. De groep items ank77 bevat items voor M7 en E7, die dus zowel medio groep 7 als eind groep 7 zijn afgenomen en de groep items ank78 bevat items voor E7 en M8. Een item kan hoogstens in één (overlap)groep voorkomen, dat wil zeggen: de ank-blokjes hebben geen gemeenschappelijke items en ook geen gemeenschappelijke items met de reguliere blokjes E3, M4, E4, M5, E5, M6, E6, M7, E7 en M8.

Longitudinale opzet
Een volledig longitudinaal design impliceert dat een cohort leerlingen gevolgd wordt van M3 tot en met M8. Een dergelijk design heeft een aantal zwaarwegende nadelen. Het is onvermijdelijk dat er uitval zal plaatsvinden. Bij ernstige selectieve uitval wordt het steeds ingewikkelder om betrouwbare normen op te stellen. Bovendien is een longitudinale studie belastend voor de deelnemende scholen en leerlingen. Dit brengt het risico mee van ongewenste en moeilijk controleerbare neveneffecten. Daarom is ervoor gekozen het longitudinale karakter van het onderzoek in te perken, en aan de deelnemende scholen te vragen deel te nemen op drie opeenvolgende meetmomenten, waarbij het startmoment verspreid is voor verschillende scholen. Bijvoorbeeld: school A start met groep 4 op het eindmoment van schooljaar x en zal eveneens deelnemen aan de opvolgende momenten M5 (schooljaar x+1) en E5 (schooljaar x+1). School B zal starten op moment M5 (schooljaar y) en zal eveneens deelnemen aan de opvolgende momenten E5 (schooljaar y) en M6 (schooljaar y+1). Op deze manier wordt rekening gehouden met de belasting voor scholen en worden toch de benodigde longitudinale data verkregen.
Aansluitend bij de verbondenheid van het design via opeenvolgende toetsmomenten en de longitudinale opzet is de kalibratie per leerjaar uitgevoerd op een beperkt deel van de gemeenschappelijke schaal. De kalibratie vond steeds plaats op basis van de verzamelde data voor dat leerjaar op de afname-momenten, aangevuld met de gegevens van het voorgaande en het opvolgende afnamemoment voor de groepen 4 t/m 7. Voor groep 8 werd aangevuld met gegevens van enkel het voorgaande afnamemoment omdat er geen volgend afnamemoment is (er is geen E-moment in groep 8). In het geval van leerjaar 5 met afnamemomenten M5 en E5 vond de kalibratie plaats op basis van afnamemomenten E4, M5, E5 en M6. Dit sluit aan bij de inhoudelijke kenmerken van de aangeboden opgaven, een sterke leerling in groep 5 zal wel opgaven uit groep 6 kunnen maken, maar geen opgaven uit groep 8 omdat deze qua inhoud nog niet allemaal zijn behandeld. Op deze manier kon dus beter rekening gehouden worden met de uitbreidingen in het onderwijsaanbod. In het geval van groep 8 vond de kalibratie plaats op basis van afnamemomenten E7 en M8.
Voor kalibratie en normering van de toetsen van elke jaargroep is op een gedeelte van het eerder vermelde design gefocust. In het geval van groep 8 betreft het dus het gedeelte van het macrodesign dat in figuur 4.2 hieronder is weergegeven.

Figuur 4.2 Gedeelte macrodesign waarop kalibratie leerjaar 8 is gebaseerd

	E7		M8
Juni 2014	E7	Ank78	
Januari 2015		Ank78	M8

Opgemerkt dient te worden dat de normering onafhankelijk is van de aangeboden items, mits deze qua inhoud passen bij de jaargroep en passen op de kalibratieschaal. De normering wordt immers gebaseerd op de vaardigheid op dat afnamemoment. De afgenomen toets is slechts een middel om de vaardigheid te bepalen. De opzet van de kalibratie en de normering zullen in de volgende paragrafen verder worden beschreven.
Om de prestaties van leerlingen en groepen te kunnen blijven volgen, worden deze op een overkoepelende schaal geplaatst door gebruik te maken van een transformatie. Deze transformatie wordt afgeleid uit de overlappende populaties op de kalibraties. De overlappende jaargroepen op opvolgende schalen bestaan uit dezelfde leerlingen in beide kalibraties en hebben per definitie dezelfde vaardigheidsverdeling. Om deze reden kan uit de vaardigheidsverdelingen van die jaargroepen de transformatie berekend worden.

4.2 Opzet en verloop van het kalibratie- en normeringsonderzoek

Met het oog op het ontwikkelen van de toetsen Begrijpend luisteren groep 8 zijn in 2013 en 2014 opgaven geconstrueerd. In 2015 zijn deze opgaven in een kalibratieonderzoek (proefonderzoek) voorgelegd aan leerlingen van groep 8 op een groot aantal scholen om gegevens te verzamelen over de kwaliteit en de moeilijkheid van de opgaven. Aansluitend zijn bij een landelijke normgroep referentiegegevens verzameld door de psychometrisch en inhoudelijk meest geschikte opgaven voor te leggen aan leerlingen op het normeringsmoment medio schooljaar (M-moment). De normering voor het M-moment vond plaats in januari en begin februari 2016. Er was geen normering op het E-moment: in verband met eindejaarsactiviteiten voor de schoolverlaters in mei/juni vindt in jaargroep 8 geen toetsing op het E-moment plaats.

Het kalibratieonderzoek
Het kalibratieonderzoek levert gegevens op over de kwaliteit en de moeilijkheid van de opgaven. In het kalibratieonderzoek, dat aan de opgavenbanken ten grondslag ligt, is uitgegaan van een onvolledig maar 'verbonden' design, zoals beschreven in paragraaf 4.1: niet alle leerlingen in de steekproef van het kalibratieonderzoek maakten alle opgaven. Opgaven werden verdeeld over taken en aan elke leerling werden meerdere taken voorgelegd. De taken die gezamenlijk aan een groep leerlingen worden voorgelegd, worden 'boekjes' ('booklets') genoemd. De verschillende boekjes overlappen elkaar. Deze overlap zorgt ervoor dat het design verbonden is, een noodzakelijke voorwaarde om CML-schattingen van de itemparameters te kunnen bepalen.

Voor een proef- annex kalibratieonderzoek zoals dit, is het niet nodig om hoge eisen te stellen aan de representativiteit van de steekproef. Niettemin werden scholen geworven op basis van eenzelfde steekproefkader als hieronder is beschreven voor het normeringsonderzoek, waarmee spreiding werd nagestreefd naar regio, schoolgrootte en type school (in termen van gewichtsleerlingen). Op deze manier hielden we bij de inrichting van het proefonderzoek al rekening met de representativiteit die in een latere fase (bij het normeringsonderzoek) wél vereist is. Een evaluatie van de representativiteit van de steekproef voor het proefonderzoek zoals we die in paragraaf 4.3 uitvoerig rapporteren voor het normeringsonderzoek laten we hier achterwege. In het kalibratieonderzoek van januari 2015 zijn 190 items voorgelegd aan 1848 leerlingen van groep 8. De 190 items en 20 ankeritems van groep 8 waren verdeeld over 9 verschillende opgavenboekjes in een onvolledig, maar 'verbonden' design. Elk boekje bestond uit ongeveer 45 items. Elk item kwam in twee boekjes voor. Het gemiddeld aantal leerlingantwoorden per item was 395.
Scholen uit de getrokken steekproef werden via een brief voor dit onderzoek uitgenodigd.
Het onderzoek verliep op de volgende wijze: de scholen ontvingen opgavenboekjes, bijbehorende antwoordbladen, een dvd met daarop de luisterfragmenten en de vragen en een handleiding voor de docent. De toets werd afgenomen door de leerkracht aan de hand van de handleiding, in overeenstemming met de situatie zoals die van toepassing is bij de uitgegeven toets. De ingevulde antwoordbladen werden door Cito verwerkt en de scholen ontvingen een rapportage met toetsscore en vaardigheidsniveau van de leerling.

Tabel 4.1 Afnamedesign kalibratieonderzoek (proefonderzoek) groep 8 (januari 2015)

Bk	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
1	E7	M8	M8											
2	M8			M8	M8									
3			M8			M8	M8							
4		M8				M8		M8						
5								M8	M8	M8				
6				M8					M8		M8			
7					M8							E7	M8	
8							M8			M8				E7
9											M8		M8	

Op grond van de gegevens uit het kalibratieonderzoek is een selectie van items gemaakt voor het normeringsonderzoek van de toets M8 voor het afnamemoment M8.

4.3 Samenstelling van de normeringssteekproef en representativiteit

Voor het normeringsonderzoek van de toets M8 waren circa 1000 leerlingen nodig. Voor het totaal aantal scholen in Nederland (6703 scholen) is een indeling gemaakt naar LVS-strata (1 t/m 5 procent, 6 t/m 10 procent, 11 t/m 20 procent en 21 procent of meer gewichtsleerlingen) bij schoolgrootte (meer dan 200 leerlingen dan wel minder dan 200 leerlingen). Dit resulteerde in 8 groepen. Vervolgens zijn clustersteekproeven getrokken op dusdanige wijze dat de 8 groepen representatief waren vertegenwoordigd in de steekproef. Bij de steekproeftrekking werd de inschatting van deelnamebereidheid gebaseerd op voorgaande wervingen. Uit deze wervingen bleek dat het gemiddelde aantal deelnemende leerlingen per school 24 was en de deelnamebereidheid aan normeringsonderzoeken 8%.

De normeringsgroep voor M8 bestond deels uit herhalingsscholen die ook op het afnamemoment E7 hadden meegedaan aan het normeringsonderzoek (‘herhalingsschool’ impliceert hier dat de betreffende scholen zich bereid verklaard hadden om meer dan één keer deel te nemen). Deze scholen waren voor de normering E7 geworven via dezelfde indeling naar LVS-strata zoals die hierboven is beschreven. Zie voor een evaluatie van de representativiteit van de normeringssteekproef E7 de wetenschappelijke verantwoording van de toets Begrijpend luisteren groep 7 (Van Berkel e.a., 2016). Voor het normeringsonderzoek M8 zijn uiteindelijk in totaal 54 herhalingsscholen van E7 aangeschreven en 56 extra scholen. Van de herhalingsscholen uit het normeringsonderzoek E7 bleken er 53 bereid deel te nemen aan het normeringsonderzoek M8. Van de scholen uit de aanvullende steekproef bleken er 4 bereid deel te nemen aan het normeringsonderzoek. Uiteindelijk resulteerde dit in een deelnamebereidheid van zo’n 98% van de herhalingsscholen en 7% van de overige aangeschreven scholen. In totaal meldden zich 57 scholen aan voor het normeringsonderzoek M8 met in totaal 1484 leerlingen. Uiteindelijk hebben 41 scholen met 944 leerlingen daadwerkelijk meegedaan en konden we de gegevens van 940 leerlingen gebruiken. Leerlingen met meer dan 25% missings van de antwoorden werden namelijk niet meegenomen bij de kalibratie en normering. Dit betreft bijvoorbeeld leerlingen die bijvoorbeeld deel 2 en 3 van de toets niet hebben gemaakt wegens ziekte.

Representativiteit van de normgroep M8
De representativiteit van de steekproeven voor het normeringsonderzoek M8 is geëvalueerd naar verdeling over het percentage leerlingen op school met een afwijkend leerlinggewicht (dus 0,3 en 1,2;), de schoolgrootte, regio, verstedelijking en sekse.

Percentage leerlingen met een afwijkend leerlingewicht
Cito maakt bij de steekproeftrekking gebruik van informatie over de percentages leerlingen met een afwijkend leerlinggewicht. Dit percentage kan tot op zekere hoogte worden opgevat als een indicatie voor het aantal achterstandsleerlingen van een school. In tabel 4.2 wordt de gewichtenregeling uitgelegd.

Tabel 4.2 Gewichtenregeling

Gewicht	Uitleg
0,0	Leerling van wie één van de ouders of beide ouders een opleiding heeft gehad van minimaal 3 jaar vmbo gemengde leerweg/vmbo theoretische leerweg/mavo, minimaal 2 jaar havo/vwo, mbo, hbo of universiteit.
0,3	Leerlingen van wie beide ouders of de ouder die belast is met de dagelijkse verzorging een opleiding heeft gehad van maximaal lbo/vbo, praktijkonderwijs of vmbo basis- of kaderberoepsgerichte leerweg.
1,2	Leerlingen van wie beide ouders een opleiding hebben gehad van maximaal basisonderwijs of (v)so-zmlk of van wie één van de ouders een opleiding heeft gehad van maximaal basisonderwijs of (v)so-zmlk en de andere ouder maximaal lbo/vbo praktijkonderwijs of vmbo basis- of kaderberoepsgerichte leerweg.

Voor elke school is het percentage leerlingen ('p') met een gewicht afwijkend van 0,0 bepaald. Gebaseerd op deze gewichten zijn er vier groepen scholen gevormd. De CFI-gegevens van oktober 2014 zijn als basis voor het steekproefkader voor de normering van M8 genomen. De verdeling van de scholen is weergegeven in tabel 4.3.

Tabel 4.3 Scholen uit steekproef M8 naar % leerlingen met afwijkend leerlinggewicht

% leerlingen met afwijkend leerlingewicht	Landelijk	Steekproef M8	
	%	Aantal	%
0% ≤ p < 10%	63,5	26	63,4
10% ≤ p < 25%	24,7	7	17,1
25% ≤ p < 40%	7,5	5	12,2
40% ≤ p < 100%	4,3	3	7,3
Totaal	100	41	100

In de steekproef van normeringsmoment M8 is het stratum 2 (10% ≤ p <25%) licht ondervertegenwoordigd en zijn stratum 3 (25% ≤ p < 40%) en 4 (40% ≤ p < 100%) licht oververtegenwoordigd. Deze verschillen zijn echter niet significant (chi-kwadraat = 3,038, df =3 en p = 0,39). Aangenomen wordt daarom dat de verzameling scholen in de steekproef representatief is naar stratum.

Schoolgrootte
Een volgende controle is naar schoolgrootte: klein en groot. Een kleine school telt minder dan 200 leerlingen; een grote school 200 of meer leerlingen. Deze verschillen worden weergegeven in tabel 4.4. Door de steekproef naar schoolgrootte te trekken wordt voorkomen dat er bijvoorbeeld alleen maar hele grote scholen in de steekproef terechtkomen, die mogelijk gemiddeld als school anders zouden kunnen presteren dan kleine scholen. Als enkele (grote) scholen uiteindelijk niet meedoen aan het onderzoek, bestaat bovendien het risico dat er te weinig data terugkomen.

Tabel 4.4 Scholen uit steekproef M8 naar schoolgrootte

Schoolgrootte	**Landelijk**	**Steekproef M8**	
	%	**Aantal**	**%**
klein	52,1	25	61,0
groot	47,9	16	39,0
Totaal	100	41	100

Voor de steekproef M8 zijn de verschillen niet significant (chi-kwadraat = 1,294, df = 1 en p = 0,26). Aangenomen wordt daarom dat de scholen in de steekproef representatief zijn naar schoolgrootte.

Regio
Bij de definitie van de variabele *regio* is uitgegaan van de CBS-indeling naar landsdeel. Dit betekent dat er vier regio's onderscheiden zijn. Regio *noord* omvat de provincies Groningen, Friesland en Drenthe; regio *oost* de provincies Overijssel, Gelderland en Flevoland; regio *west* de provincies Utrecht, Noord-Holland, Zuid-Holland en Zeeland en regio *zuid* de provincies Noord-Brabant en Limburg.

De verdeling van alle scholen in de populatie en de scholen in de steekproef van groep 8 naar regio staat in tabel 4.5. In de steekproef van normeringsmomenten M8 zijn de scholen in regio oost wat ondervertegenwoordigd en de scholen in regio zuid en noord wat oververtegenwoordigd. Dit verschil is echter niet statistisch significant (chi-kwadraat = 3,199, df = 3, p = 0,36). Aangenomen wordt daarom dat voor groep 8 de scholen representatief zijn naar regio.

Tabel 4.5 Scholen uit steekproef M8 naar regio

Regio	**Landelijk**	**Steekproef M8**	
	%	**Aantal**	**%**
west	41,7	15	36,6
oost	24,6	7	17,1
noord	15,2	9	22,0
zuid	18,5	10	24,4
Totaal	100	41	100

Verstedelijking

De populatieverdelingen van de scholen en de verdeling van de scholen in de steekproef naar verstedelijking (urbanisatiegraad) staan in tabel 4.6. Het betreft hier een indeling in vijf categorieën die bij het CBS gebruikelijk is. In de steekproef M8 zijn de scholen in matig en weinig verstedelijkte gebieden wat oververtegenwoordigd en scholen in zeer sterk en sterk verstedelijkte gebieden wat ondervertegenwoordigd. Het verschil is niet significant (chi-kwadraat = 5,614, df = 4 en p = 0,23). Aangenomen wordt daarom dat voor groep 8 de scholen in de steekproef nagenoeg representatief zijn naar verstedelijking.

Tabel 4.6 Scholen uit steekproef M8 naar mate van verstedelijking

Mate van verstedelijking	**Landelijk**	**Steekproef M8**	
	%	**Aantal**	**%**
niet	19,7	8	19,5
weinig	19,4	13	31,7
matig	19,7	10	24,4
sterk	22,4	8	19,5
zeer sterk	12,2	2	4,9
Totaal	100	41	100

Sekse

Bij de variabele *sekse* is een tweedeling naar jongens en meisjes gehanteerd. De verdeling van alle leerlingen en de leerlingen in de steekproef van groep 8 naar sekse staat in tabel 4.7. In de steekproef zijn de jongens enigszins ondervertegenwoordigd. Echter ook hier is geen significant verschil geconstateerd (chi-kwadraat = 3,457; df = 1 en p = 0,063).

Tabel 4.7 Leerlingen uit steekproef M8 naar sekse

Sekse	**Landelijk**	**Steekproef M8**	
	%	**Aantal**	**%**
jongen	50,4	393	47,2
meisje	49,6	440	52,8
Totaal	100	833	100

De deelnemende scholen zijn dus voor alle achtergrondvariabelen representatief te noemen voor de populatie van scholen. Statistische weging is om die reden dan ook niet nodig. Hetzelfde geldt op leerlingniveau voor de verdeling naar sekse.

Het is niet mogelijk om expliciet rekening te houden met de variabele *etniciteit*, omdat er geen eenduidige referentiegegevens voor de populatie bekend zijn. Aan scholen die meedoen aan het normeringsonderzoek is wel gevraagd in leerlinglijsten met achtergrondgegevens ‘thuistaal’ in te vullen, maar deze gegevens werden door de meeste scholen niet verstrekt.

Eerder peilingsonderzoek heeft echter laten zien dat de verdeling naar etnische herkomst sterk samenhangt met de verdeling naar urbanisatiegraad en percentages gewichtenleerlingen (Hemker, Kordes en Van Weerden, 2011). Om deze reden is aangenomen dat de uiteindelijke normeringsteekproef voldoende representatief is naar etnische herkomst als de verdeling naar urbanisatiegraad en percentages gewichtenleerlingen overeenkomt met de verdeling in de landelijke populatie.

4.4 Kalibratie

4.4.1 De kalibratieprocedure

Met kalibratie wordt bedoeld dat we bij de items kengetallen zoeken die de antwoorden van de leerlingen goed representeren. Hoe de kengetallen gezocht worden, ligt deels vast door het gekozen model (zie paragraaf 2.4.2.2). Hoe succesvol deze operatie is, kan statistisch getoetst worden. Eenvoudig gezegd, schatten we in OPLM met de CML-methode de itemparameters en controleren we of deze de data goed voorspellen. Voor een exacte beschrijving van de statistische toetsen die in OPLM gebruikt worden, hun eigenschappen en feitelijke implementatie in OPLM, verwijzen we naar Verhelst (1993). Hier beperken we ons tot een korte beschrijving van de principes van de statistische toetsen die gebruikt zijn in de kalibratie-procedure. De statistische toetsen in OPLM hebben goede statistische en asymptotische eigenschappen, daar OPLM behoort tot de exponentiële familie, met de gewogen somscore:

$$s = \sum_{i=1}^{k} a_i x_i \,, \tag{4.1}$$

als een 'afdoende statistiek' voor de vaardigheid θ. Dit betekent dat alle informatie in de data met betrekking tot de vaardigheid in deze statistiek aanwezig is. Hiervan wordt gebruikgemaakt bij de statistische toetsen in OPLM. Het basisprincipe van de statistische toetsen in OPLM is dat op grond van de afdoende statistiek s de personen in de data kunnen worden gegroepeerd. En binnen deze groepen kan de verwachte proportie goede antwoorden op een item onder het model, $p(+\mid s)$, vergeleken worden met de feitelijk geobserveerde proportie goede antwoorden, $prop(+\mid s)$. In het polytome geval worden de items gedichotomiseerd, de proportie goede antwoorden verwijst dan naar de hoge itemscore (zie Verhelst, 1993, hoofdstuk 7). Via de basisvergelijking van OPLM kunnen we eenvoudig de conditionele kans op het goed beantwoorden van de items afleiden en daarmee kunnen we $p(+\mid s)$ evalueren, $prop(+\mid s)$ volgt uit de data. Discrepanties tussen $p(+\mid s)$ en $prop(+\mid s)$ duiden op schendingen van het model.
Deze discrepanties vormen de basis voor de diverse statistische toetsen in OPLM. De toetsingsgrootheid voor de veronderstelde discriminatie-indices is gegeven door

$$M = f_{s\in H}(p(+\mid s) - prop(+\mid s)) + f_{s\in L}(prop(+\mid s) - p(+\mid s)). \tag{4.2}$$

Deze zogeheten M-toetsen verdelen de scoregroepen in een laag deel (L) en een hoog deel (H) en f is een monotone functie. M-toetsen hebben een duidelijke interpretatie: is M significant positief dan is de veronderstelde steilheid van de ICC (item karakteristieke curve) overschat in het model, is M daarentegen erg laag dan is de index te klein. Verhelst laat zien voor welke functie, f , $M \approx N(0,1)$. In OPLM zijn drie verschillende M-toetsen geïmplementeerd die verschillen in de definitie van de hoge en lage scoregroepen. Naast deze M-toetsen is er een algemene itemtoets die de volgende vorm heeft

$$S = f(p(+\mid s) - prop(+\mid s)) \,.$$

Deze zogenoemde S-toets heeft een χ^2 verdeling onder het model. Als globale modeltoets is de R1c-toets (Glas, 1988) geschikt. Ook de distributie van alle afzonderlijke S-toetsen komt hiervoor in aanmerking. Als we deze S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn op het (0,1) interval.
Kortom, als we afzien van de formeel-statistische achtergrond van de gehanteerde toetsen, kan de kalibratieprocedure als volgt worden samengevat:

1. Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
2. Vervolgens schatten we de itemparameters met behulp van de CML-methode.
3. Met behulp van de M-toetsen controleren we of de discriminatie-indices goed zijn ingesteld.

4. Een volgende controle betreft de overschrijdingskansen van de S-toetsen en een grafische modelcontrole door middel van het programma Wopplot (grafische inspectie van de ICC's).
5. Vervolgens vindt een globale modelcontrole plaats in de vorm van een R1c-toets en de verdeling van de overschrijdingskansen van de S-toetsen.

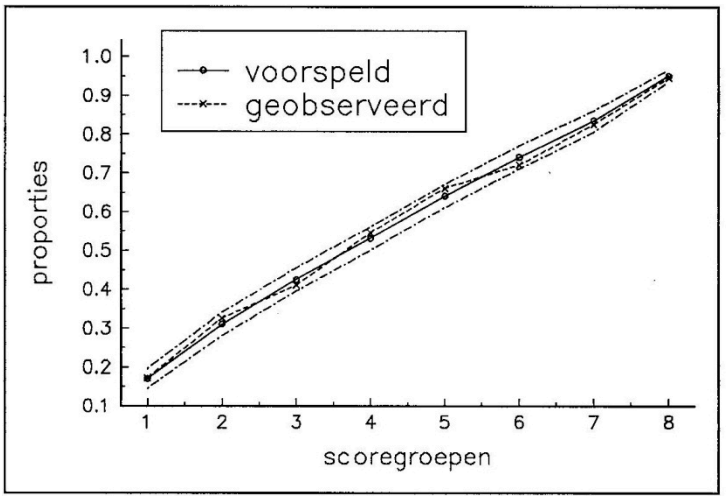
De stappen 1 tot en met 5 worden een aantal malen doorlopen tot het resultaat bevredigend is. Afhankelijk van de uitkomsten kunnen items worden verwijderd. De opgaven vormen na de kalibratie een gekalibreerde opgavenbank, waarbij de opgaven per onderscheiden vaardigheidsdimensie een beroep doen op hetzelfde complex aan vaardigheden of 'latente trek'.

OPCAT voert een aantal statistische toetsen uit op grond waarvan bepaald kan worden of het model een adequate beschrijving geeft van de data. Belangrijk zijn de zogenaamde itemgeoriënteerde *S*-toets en de overall R_{1c}-toets. De *S*-toets is asymptotisch χ^2 verdeeld en is gebaseerd op de verschillen tussen de geobserveerde en verwachte proporties antwoorden in homogene scoregroepen. Een uniforme verdeling van *p*-waarden voor de *S*-toetsen in het interval [0,1] pleit voor passing van het model. De R_{1c}-toets heeft dezelfde onderliggende rationale als de *S*-toets en wordt over het algemeen acceptabel bevonden indien zijn waarde niet groter is dan anderhalf tot hooguit twee keer het aantal vrijheidsgraden.

4.4.2 Resultaten van de kalibratieprocedure: modelfit

Het is niet eenvoudig om de kwaliteit van de kalibratie aan te tonen. De belangrijkste statistische instrumenten om de passing van een opgave in het IRT-model te bewerkstellingen en uiteindelijk te documenteren betreffen de hierboven al besproken S-toetsen. Het lastige daarvan is, dat de toetsing voor een groot deel visueel gebeurt. Dit kunnen we illustreren aan de hand van figuur 4.3 (zie Staphorsius, 1994, blz. 239). Ten behoeve van deze toetsing wordt de totale groep van leerlingen die een verzameling opgaven gemaakt heeft, ingedeeld in een aantal (meestal acht) scoregroepen. Elke groep bestaat uit leerlingen met een ongeveer even hoge score. De geobserveerde proporties juiste antwoorden van deze groepen (telkens gesymboliseerd door een x) zijn door de middelste stippellijn verbonden. De volle lijn daarentegen verbindt de proporties die op grond van de parameterschattingen voorspeld kunnen worden. De twee buitenste lijnen geven het 95%-betrouwbaarheidsinterval aan. De breedte van dit interval is in belangrijke mate afhankelijk van het aantal leerlingen dat de opgave heeft beantwoord.
Uit de figuur blijkt heel duidelijk dat de geobserveerde proporties, zoals bedoeld, binnen het 95%-betrouwbaarheidsinterval van de (geschatte) voorspelde proporties liggen, en dit komt in grote lijnen overeen met een niet-significante S-toetsingsgrootheid (Verhelst et al., 1994).

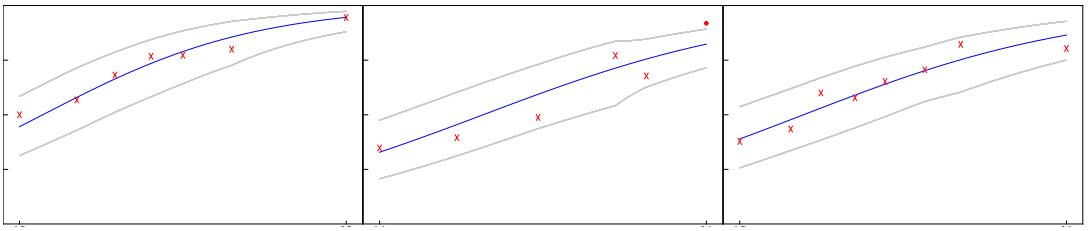
Figuur 4.3 Grafische voorstelling van een S_i–toets



Het is ondoenlijk om voor alle opgaven dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.4 illustreren dat zelfs bij de slechtst passende opgave sprake is van een zeer aanvaardbaar beeld. Er wordt in deze gevallen voor een deel (van de onderscheiden scoregroepen) niet beantwoord aan de eis dat de geobserveerde proportie binnen het 95%-betrouwbaarheidsinterval van de geschatte proporties ligt. Dit beeld doet zich slechts bij enkele opgaven voor die dan ook een uitzondering vormen. De overige opgaven voldoen voor alle scoregroepen wel aan die eis. De afbeeldingen voor de representatieve en best passende opgave illustreren dit. Dit leidt tot de conclusie dat bij vrijwel alle opgaven in de toetsen Begrijpend luisteren een grafische voorstelling van de S–toetsing hoort die in grote lijnen met figuur 4.3 overeenkomt; andere opgaven zijn bij de kalibratie niet in de itembank opgenomen. Dit is, zeker gezien de relatief grote aantallen observaties die in het geding zijn, een zeer sterke aanwijzing dat het meetinstrument en het meetmodel dat ontwikkeld is, respectievelijk gebruikt is, adequaat zijn om het gedrag van de leerlingen te verklaren. Bovendien blijkt, en dat is vanuit theoretisch oogpunt nog belangrijker, dat gemeten verschillen in gedrag tussen de leerlingen te verklaren zijn door één unidimensioneel concept.

Figuur 4.4 Voorbeelden van S-toetsen voor de toets Begrijpend luisteren M8 met de best passende, de slechtst passende en een qua passing representatieve opgave

Best passend Slechtst passend Representatieve passing



In feite kan men bij de kalibratie beter varen op deze grafische weergaven dan op toetsingsresultaten in termen van exacte getallen en de significantie daarvan. Niettemin zijn er bij de kalibratie S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij zijn we vooral geïnteresseerd in de distributie van de overschrijdingskansen van deze verzameling toetsingsresultaten. Zoals eerder aangegeven zouden de overschrijdingskansen gelijkmatig verdeeld moeten zijn binnen het (0,1) interval, uiteraard met zo weinig mogelijk significante resultaten. Tabel 4.8 waarin het (0,1) interval is opgedeeld in tien gelijke stukken, geeft een beeld van de uitkomsten bij een kalibratie van alle 33 opgaven van de toets Begrijpend luisteren groep 8. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan 0,01, respectievelijk 0,05. Het is duidelijk dat voor de toets de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Deze resultaten geven een bevestiging van het eerder geschetste beeld, dat met uitzondering van enkele opgaven (drie opgaven met een p-waarde < 0,05; geen enkele opgave heeft een p-waarde < 0.01), sprake is van niet-significante S-toetsen. Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensioneel construct representeren.

Tabel 4.8 Verdeling van overschrijdingskansen bij S-toetsen voor M8

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1											
M8	0	3	1	5	1	3	2	0	5	5	3	5

In tabel 4.9 is de R1c-waarde weergegeven voor de toets Begrijpend luisteren M8 waarvoor in tabel 4.8 de resultaten van de S-toetsen (op itemniveau) zijn weergegeven. R1c is een statistiek die zicht geeft op de

modelpassing van de toets als geheel. Voor een acceptabele modelfit geldt als vuistregel dat R1c niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden (df) zou mogen zijn en bij voorkeur niet-significant.
De modelpassing van de toets voldoet aan een van deze voorwaarden. Voor M8 geldt dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootheid is significant. Aan dit laatste moet bij steekproeven met een dergelijke omvang (940 leerlingen in totaal) niet te veel waarde worden gehecht.

Tabel 4.9 R1c-waarde voor M8

Toetsversie	R1c	df	p
M8	287,333	232	0,007

Ten slotte bespreken we nog een methode om de modelpassing te verantwoorden die wordt besproken in het COTAN Beoordelingssysteem (Evers et al., 2010). Het betreft hier een poging om de nauwkeurigheid van de itemparameterschattingen te beoordelen op basis van een constante (in het COTAN Beoordelingssysteem met 'c' aangeduid) die weergeeft hoe de relatie is tussen de standaardfout van de moeilijkheidsparameter van een item en de geschatte standaarddeviatie van de vaardigheidsverdeling in de populatie. Het beoordelingssysteem geeft ook richtlijnen voor het beoordelen van de grootte van deze 'c'. Deze dient te worden beoordeeld als goed als de waarde lager is dan of gelijk aan 0,20. Waarden tussen 0,30 en 0,40 kunnen nog als voldoende worden beschouwd. De standaardfouten van de moeilijkheids-parameters worden dus gedeeld door de standaarddeviatie van de populatie waarin ze zijn afgenomen. Voor geen enkele opgave is de waarde groter dan 0,30. Range en gemiddelde duiden op een hoge nauwkeurigheid van de itemparameterschattingen. Zie tabel 4.10.

Tabel 4.10 Nauwkeurigheid van de itemparameterschattingen (constante 'c')

Toetsmoment	Constante 'c'	
	Range	Gemiddelde
M8	0,133 – 0,262	0,196

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toets Begrijpend luisteren van het Cito Volgsysteem primair en speciaal onderwijs (LVS) voor de afnamemoment M8 de kalibratie geslaagd is. Hiermee is het laatste woord nog niet gezegd over de validiteit, maar het kalibratie-onderzoek brengt in ieder geval een essentieel aspect van het validiteitsvraagstuk naar voren: de rechtvaardiging van wat in de meeste toetstoepassingen gebruikelijk is, namelijk het reduceren van alles wat de leerling heeft geantwoord tot een enkele toetsscore (of afgeleid daarvan, een enkele schatting van zijn onderliggende vaardigheid). De kalibratie-analyse, als puur formeel proces, kan geen uitspraken doen over de inhoudsvaliditeit of over de constructvaliditeit als antwoord op de vraag: hoe kan worden aangetoond dat het concept dat de items in de bank meten dekkend is voor en samenvalt met het construct dat we in de toetsen Begrijpend luisteren proberen te meten (zoals dat in het didactisch en het wetenschappelijk forum wordt bedoeld)? In hoofdstuk 6 over validiteit zal worden nagegaan of het gemeten concept inderdaad overeenkomt met het begrip zoals bedoeld. De vraag is dan in het geval van de toets Begrijpend luisteren: kan het unidimensionele concept onder de opgaven in de opgavenbank Begrijpend luisteren inderdaad worden opgevat als de vaardigheid 'begrijpend luisteren'? Een geslaagde kalibratie op een unidimensioneel construct beschouwen we als een noodzakelijke voorwaarde voor deze begripsvaliditeit.

4.5 Normeringsresultaten

Aan het normeringsonderzoek M8 namen 940 leerlingen deel. In paragraaf 2.4.2 gaven we belangrijke implicaties voor een gekalibreerde opgavenverzameling. Het slagen van de kalibratie betekent dat we met een selectie van opgaven uit de opgavenbank de vaardigheid bij een leerling kunnen meten.
Hoe nauwkeurig we dat doen, is beschreven in paragraaf 5.2.
We kunnen nu een schatting maken van de verdelingen van de vaardigheid in de populatie leerlingen van groep 8 op afnamemoment M8, omdat we de toetsopgaven voorgelegd hebben aan een aselecte steekproef van leerlingen die in overeenstemming is met deze afnameperiode. We schatten het gemiddelde en de standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met behulp van deze gegevens kunnen dan ook schattingen gemaakt worden van de percentielen in de populatie, die van belang zijn voor de indeling van leerlingen in de niveaugroepen die zijn beschreven in paragraaf 2.3.

Deze percentielen zijn voor afnamemoment M8 weergegeven in tabel 4.11. Een overzicht van de geschatte gemiddelden en de standaardafwijkingen van de vaardigheid op dit normeringsmoment voor de onderzochte populatie is eveneens te vinden in tabel 4.11.

Tabel 4.11 Overzicht van de vaardigheidsverdeling voor het normeringsmoment en de percentielen

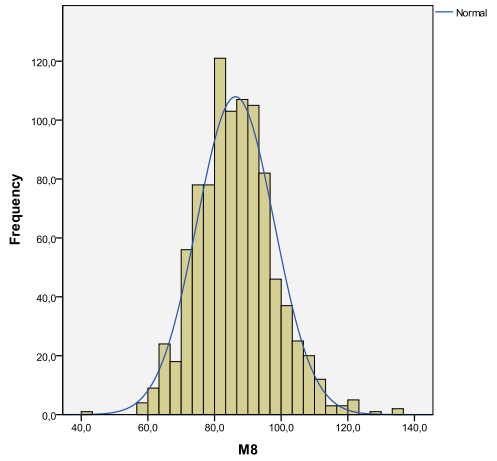
Afnamemoment	N	Gem	SD	P10	P20	P25	P40	P50	P60	P75	P80	P90
M8	940	86,4	9,77	73,9	78,2	79,8	83,9	86,45	88,9	93,0	94,6	98,9

In figuur 4.5 is in een histogram de verdeling van vaardigheidsscores voor de normeringssteekproef weergegeven, additoneel is ook de bijbehorende normaalverdeling ingetekend. De aanname van een normaal verdeelde vaardigheidsverdeling wordt ondersteund door de data. Bovendien is hiervoor ook een statistische toets ontwikkeld, de zogeheten R0 toets (Verhelst, Glas & Verstralen, 1995). Deze is niet significant (zie tabel 4.12). Dit impliceert dat de vaardigheid op het normeringsmoment M8 als normaal verdeeld kan worden opgevat.

Tabel 4.12 Toets op normaliteit van de vaardigheidsverdeling op normeringsmoment M8

Moment	R0-statistiek		
	R0	df	prob (R0)
M8	152,4	134	0,133

Figuur 4.5 Histogram van de vaardigheidsscores van de normeringssteekproef M8 inclusief normaalverdeling



Geldigheid van de normen

De toetsen van het Cito Volgsysteem primair en speciaal onderwijs worden in principe elke acht tot tien jaar vernieuwd. Niet alleen de inhoud wordt vernieuwd en aangepast aan de ontwikkelingen in het onderwijs, ook de normen worden opnieuw vastgesteld. Omdat er enige tijd verloopt tussen de dataverzameling in het normeringsonderzoek en het moment waarop de vernieuwde toets wordt uitgebracht, kan voor de toets Begrijpend luisteren groep 8 een geldigheid aangehouden worden tot en met 2025.
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.

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Methoden om de betrouwbaarheid te bepalen

In hoofdstuk 4 is beschreven hoe de kalibratie en normering is uitgevoerd en zijn de resultaten daarvan beschreven. In dit hoofdstuk gaan we nader in op de betrouwbaarheid en de meetnauwkeurigheid van de toets M8. Het is mogelijk om de betrouwbaarheid van de toets voor elk meetmoment te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de toets OPLM-geschaald zijn.
Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het feit dat de toets volledig bestaat uit OPLM-gekalibreerde items. Om relevante beschrijvende gegevens bij de toets te genereren, is gebruikgemaakt van het programma OPLAT (Verhelst, Glas en Verstralen, 1995).

In OPLAT wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ. Deze verwachte waarde wordt aangeduid met τ(θ). Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie worden aangegeven met het symbool Var(τ). Tussen θ en τ(θ) bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de toetsscore τ(θ) moet behalen (dat is alleen zo als de toets oneindig lang wordt). De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor met een eenmalige toetsafname niet meer zonder fout de waarde van θ bepaald kan worden. De variantie van de geobserveerde toetsscore wordt aangegeven met Var(t|τ(θ)), en door weer gebruik te maken van de distributie van θ in de populatie kan ook de gemiddelde variantie van de geobserveerde toetsscores berekend worden.

$$Var(t) = E[Var(t \mid \tau(\theta))] \tag{5.1}$$

Deze variantie kan opgevat worden als de (gemiddelde) meetfoutvariantie in de metriek van de geobserveerde scores (t). In analogie met de theorie over de betrouwbaarheid volgt dan

$$MAcc = \frac{Var(\tau)}{Var(\tau) + Var(t)} \tag{5.2}$$

waarin MAcc staat voor 'Accuracy of Measurement'.

5.2 Betrouwbaarheid: resultaten

Tabel 5.1 bevat informatie over de meeteigenschappen van de toets Begrijpend luisteren. In de tweede kolom staat de maximumscore, deze is gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. De derde kolom geeft de geschatte gemiddelde scores van de leerlingen op de toets aan. De vierde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van de toets. De vijfde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de toets op de twee afnamemomenten is.

Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen Begrijpend luisteren) geeft de COTAN (Evers et al., 2010) aan dat een betrouwbaarheids-coëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 voldoende, en een

betrouwbaarheidscoëfficiënt hoger dan 0,80 goed. Op grond van dit criterium is de betrouwbaarheid van de toets op M8 voldoende te noemen (0,77). In tabel 5.1 gaat het om de toets voor groep 8 op afnamemoment M8.

Tabel 5.1 Beschrijvende gegevens bij de toets Begrijpend luisteren M8 voor afnamemoment M8

Afnamemoment	Maximum score	Gemiddelde	Standaard-meetfout	MAcc	Test-hertest (simulatie)
M8	33	23,8	2,36	0,77	0,77

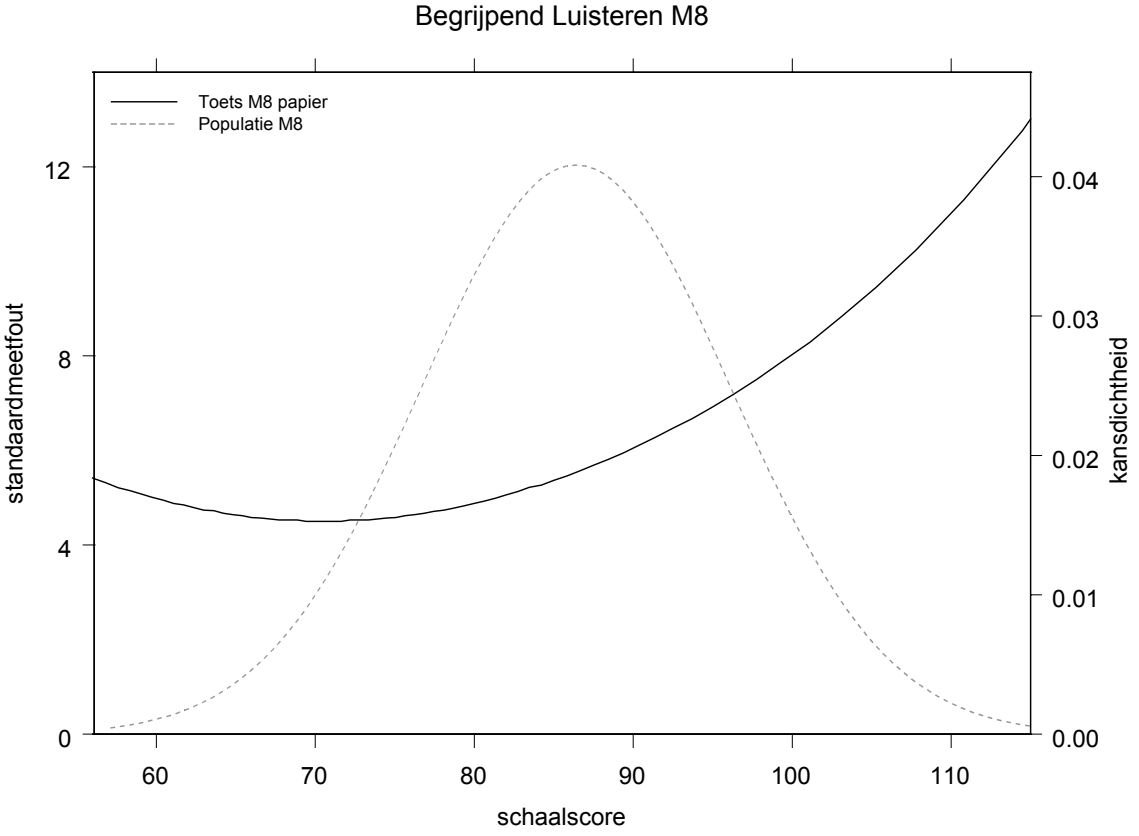
Er heeft geen test-hertest onderzoek plaatsgevonden. De afnamecontext van de LVS-toets Begrijpend luisteren leent zich daar niet goed voor. Het feit dat alle items echter OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1.000.000 leerlingen. Daarbij hebben we enerzijds de vaardigheidsverdeling van alle leerlingen, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselect uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 1.000.000 dubbele (virtuele) afnames berekend. Men kan deze simulatie beschouwen als een test-hertestonderzoek onder ideale condities. De tweede toetsafname is immers volledig onafhankelijk van de eerste toetsafname en wordt niet beïnvloed door de kennis die de leerling mogelijk verworven heeft in de eerste toetsafname. Daarnaast is er geen sprake van invloed van een test-hertest-interval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden. De resultaten zijn weergegeven in tabel 5.1.
De uitkomst komt exact overeen met eerder berekende MAcc en leidt dan ook tot dezelfde conclusie met betrekking tot de betrouwbaarheid van de toets Begrijpend luisteren.

5.3 Lokale betrouwbaarheid en meetnauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen en geven geen beeld van de lokale meetnauwkeurigheid ervan. Figuur 5.1 geeft grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de toets M8. In deze figuur staat de grootte van de meetfout op de vaardigheidsschaal afgebeeld (met verdelingskenmerken zoals aangegeven in tabel 4.11).

Ook is de kansdichtheidsfunctie voor de normgroep M8 opgenomen. Deze laat zien hoe de vaardigheid van de leerlingen verdeeld is over de vaardigheidsschaal in de groep leerlingen in de normerings-steekproef. Figuur 5.1 maakt duidelijk dat de meetfout kleiner is in de lagere en gemiddelde vaardigheidsregionen dan in de hogere vaardigheidsregionen.

Figuur 5.1 Grootte van de meetfouten voor de toets M8 en de kansdichtheidsfunctiesvoor de M8 -populatie



Betrouwbaarheidstabellen

De betekenis van de (lokale) meetnauwkeurigheid voor de beslissingen die met de toets genomen worden, is af te leiden uit een betrouwbaarheidstabel. Tabel 5.2 laat voor afnamemoment medio groep 8 zien hoe vaak de werkelijke vaardigheidsscore in dezelfde scoregroep valt als de geschatte vaardigheidsscore. Zo laat tabel 5.2 zien dat 82,9 procent van de leerlingen die halverwege groep 8 op basis van de M8-toets in scoregroep V geclassificeerd wordt ook met hun werkelijke vaardigheidsscore in deze scoregroep geclassificeerd wordt. De kans dat een V-leerling terecht als V-leerling wordt bestempeld is, met andere woorden, meer dan 80 procent. Verder laat de linkerkant van tabel 5.2 zien dat 14,9 procent van de leerlingen in scoregroep V een vaardigheidsscore heeft die in werkelijkheid in scoregroep IV valt. De overige getallen in tabellen 5.2 zijn op dezelfde wijze te interpreteren.

Tabel 5.2 Betrouwbaarheidstabel toets M8 voor afnamemoment medio 8

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	V	IV	III	II	I	Score-groepen E t/m A	E	D	C	B	A
V	82,9	14,9	2,1	0,2	0,0	E	75,9	20,7	3,3	0,1	0,0
IV	28,3	41,7	22,8	6,7	0,6	D	26,9	44,6	25,4	3,0	0,1
III	7,4	26,1	33,6	25,4	7,6	C	4,5	23,4	45,1	23,3	3,8
II	1,8	10,3	22,8	34,4	30,6	B	0,5	5,1	25,2	40,6	28,6
I	0,3	2,0	6,1	16,0	75,5	A	0,1	0,7	5,3	17,3	76,7

In de onderzoeksliteratuur is weinig geschreven over de beoordeling van betrouwbaarheidstabellen. Wanneer een betrouwbaarheidstabel als goed of voldoende kan worden beschouwd is onduidelijk en wat verwacht mag worden onder ideale omstandigheden is, voor zover ons bekend, niet onderzocht. Daarom worden betrouwbaarheidstabellen vaak samengevat in één of meerdere indices. Wij gebruiken de *plus/minus 1 niveau-index* en de *Marginal Classification Accuracy*. De eerste maat is bedacht door Pilliner (1969). Hij stelt als ambitieniveau dat 95 procent van de leerlingen in een scoregroep in werkelijkheid ook in die scoregroep moet scoren, ***of*** één scoregroep daarboven ***of*** één scoregroep daaronder. In de tabellen 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 foutieve classificaties. In dat licht is de maximale accuraatheid die op het individuele niveau bereikt kan worden plus of minus één scoregroep. De tweede maat wordt op verschillende plekken in de literatuur beschreven. De 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* rond 0,75 - 0,80 uit te komen (Keuning & Béguin, in voorbereiding). In de praktijk liggen de waarden vaak tussen 0,60 en 0,70.

De samenvattende indices voor medio groep 8 zijn te vinden in tabel 5.3. Waar de betrouwbaarheids-tabel laat zien dat er behoorlijk wat leerlingen zijn die op basis van hun geschatte vaardigheidsscore een niveaugroep te hoog of te laag geplaatst worden, maakt tabel 5.3 aannemelijk dat de uitkomsten wel redelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969). Gemiddeld gezien scoort, afhankelijk van de gekozen indeling in scoregroepen, 92 tot 95 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* ligt boven de 56 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij het afnamemoment gemiddeld gezien in ruim 50 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. Omdat zowel de *plus/minus 1 niveau-index* als de *Marginal Classification Accuracy* wat lager liggen dan wenselijk is, moet de indeling van leerlingen in scoregroepen met de nodige voorzichtigheid geïnterpreteerd worden. De toets Begrijpend luisteren M8 weet vooral de laagst en hoogst scorende leerlingen accuraat te classificeren; in het midden is de accuraatheid van de classificatie minder. Dit pas 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 de scoregroepen II en III, respectievelijk scoregroep B.

Tabel 5.3 Samenvattende indices toets M8

	Medio 8	
	scoregroepen I t/m V	scoregroepen A t/m E
Marginal classification accuracy	56,1	56,6
Accuracy plus/minus 1 niveau	91,5	95,2

Conclusie
De vaardigheidsgroei voor Begrijpend luisteren voltrekt zich langzaam in de tijd. De verschillen tussen vaardigheidsscores op achtereenvolgende meettijdstippen zijn door de bank genomen klein, ook al neemt men slechts een maal per jaar een toets af voor deze vaardigheid. (Alleen groep 8 vormt daarop een uitzondering.) Bovendien is er sprake van meetfouten. De toch al kleine verschillen in vaardigheidsgroei moeten tegen de achtergrond van die meetfouten worden geïnterpreteerd. Dit betekent dat men weliswaar uitspraken kan doen over de vaardigheidsgroei van een leerling, maar dat deze uitspraken met voorzichtigheid dienen te worden gehanteerd. Dit geldt ook wanneer men de progressie van een leerling volgt in termen van vaardigheidsniveaus of een vergelijking maakt met andere leerlingen in termen van vaardigheidsniveaus. Want ook bij de indeling in vaardigheidsniveaus speelt de nauwkeurigheid van de

toets een rol. Hoe de leerkracht hier in de praktijk mee om moet gaan wordt toegelicht in de handleiding voor de leerkracht in paragraaf 4.3.

6 Validiteit

In de volgende paragrafen zal de validiteit besproken worden aan de hand van de inhoudsvaliditeit (6.1) en begripsvaliditeit (6.2). Criteriumvaliditeit is bij de LVS-toetsen niet aan de orde.

6.1 Inhoudsvaliditeit

De inhoudsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de opgaven in een toets een welomschreven en afgebakend universum representeren van mogelijk in de toets op te nemen opgaven. De inhoudsvaliditeit van de toetsen Begrijpend luisteren wordt onder meer gegarandeerd door de wijze waarop de opgaven ontwikkeld zijn. In de inhoudsverantwoording (zie paragraaf 3.2) is al aangegeven dat de opgaven ontwikkeld zijn door ervaren leerkrachten uit het basisonderwijs en dat aan de basis van de ontwikkeling van de opgaven de indeling in vaardigheidsaspecten ligt. Deze indeling is ontwikkeld aan de hand van de visie van Cito-toetsdeskundigen op wat het construct 'begrijpend luisteren' inhoudt en is gevoed door documenten van Sijtstra (Sijtstra, 2005) en Krom (Krom e.a., 2011), de kerndoelen voor het primair onderwijs (Ministerie van OCW, 2006) en de tussendoelen Mondelinge communicatie en leerstoflijnen (TULE, 2008). Het referentiekader Taal was op het moment van ontwikkeling van deze visie nog niet verschenen. Na verschijnen van het referentiekader Taal is deze visie op het construct 'begrijpend luisteren', gelegd naast wat in het referentiekader beschreven staat onder het subdomein Luisteren. In grote lijnen komen onze uitgangspunten overeen met wat er in het referentiekader staat.
De diverse vaardigheidsaspecten zijn uiteindelijk in voldoende mate in de toetsen vertegenwoordigd, zo blijkt uit tabel 3.1 '*Aantal opgaven Begrijpen en Interpreteren in de toets Begrijpend luisteren groep 8'* (zie hoofdstuk 3.2.2). Hierbij willen we graag nogmaals aantekenen dat de vaardigheidsaspecten niet zo duidelijk van elkaar te scheiden zijn als de indeling wellicht suggereert. Ze grijpen op elkaar in, beïnvloeden elkaar en bouwen op elkaar voort. We kunnen ze dan ook niet opvatten als te isoleren aspecten en vaardigheden van het begrijpend luisteren. Het feit dat de verzamelde data laten zien dat de opgaven op één vaardigheidsschaal liggen, illustreert dit ook. We zijn er daardoor zeker van dat het om één unidimensionale vaardigheid gaat.
Een verdere aanwijzing voor de inhoudsvaliditeit is het gegeven dat een derde van de opgaven een beroep doet op de vaardigheid Begrijpen en tweederde op de vaardigheid Interpreteren. Dit sluit aan bij de geraadpleegde literatuur, waarin de indeling Begrijpen, Interpreteren en Reflecteren beschreven wordt (vgl. de Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2009a, Krom e.a., 2011 en Sijtstra, 2005). Overigens zijn in de toetsen Begrijpend luisteren zijn alleen opgaven opgenomen die een beroep doen op de vaardigheden Begrijpen en Interpreteren, omdat met een complexe vaardigheid als Reflecteren in een evaluatieve eenrichtingssituatie nog maar weinig ervaring in het basisonderwijs is opgedaan. Bovendien zijn het ook met name de vaardigheden Begrijpen en Interpreteren die (jeugdige) luisteraars toepassen tijdens het luisteren naar gesproken taal.

6.2 Begripsvaliditeit

In deze paragraaf worden resultaten met betrekking tot verschillende aspecten van begripsvaliditeit besproken. Dit zijn de volgende: unidimensionaliteit (paragraaf 6.2.1), itemkwaliteit (paragraaf 6.2.2), itembias (paragraaf 6.2.3), convergente en discriminante validiteit (6.2.4) en verschillen tussen relevante subgroepen (6.2.5).

6.2.1 Unidimensionaliteit

In hoofdstuk 4 werd beschreven dat de opgaven van de toetsen Begrijpend luisteren na de kalibratie een gekalibreerde opgavenbank vormen. Bij de analyse van de leerlingantwoorden is nagegaan of de

verschillende opgaven van elke toets een beroep doen op hetzelfde complex aan vaardigheden. Opgaven die niet voldeden aan de passingscriteria die we beschreven in paragraaf 4.4.2, zijn uit de opgavenverzameling verwijderd. Het betreft opgaven waarop werd gegokt, opgaven die onjuist geformuleerd zijn of opgaven die een slecht onderscheidend vermogen bleken te hebben.
We hebben verschillende analyses gerapporteerd met betrekking tot de passing van het onderliggende meetmodel van de toetsen, waaruit blijkt dat die passing bevredigend is. De grafische voorstellingen van de S-toetsen gaven voor de meeste opgaven een bevredigend beeld. Dat is een sterke aanwijzing dat het meetinstrument en het meetmodel dat ontwikkeld is, respectievelijk gebruikt is, adequaat zijn om het gedrag van de leerlingen te verklaren. Het blijkt dat gemeten verschillen in gedrag tussen de leerlingen te verklaren zijn door één unidimensioneel concept. Ook de verdeling van overschrijdingskansen bij de S-toetsen voor M8 gaf een bevredigend beeld. Voor de toets M8 geldt dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt: de modelpassing van de toets voldoet dus aan de voorwaarden voor een acceptabele modelfit.
Nog een methode om de modelpassing te verantwoorden betreft het beoordelen van de nauwkeurigheid van de itemparameterschattingen op basis van een constante, de ‘c’ uit het COTAN-systeem (Evers et al., 2010). Voor M8 is voor geen enkele opgave de waarde groter dan 0,30 (tabel 4.10) en de constante kan dus beoordeeld worden als ‘goed’.
Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de LVS-toets Begrijpend luisteren M8 de kalibratie geslaagd is. De geslaagde kalibratie maakt duidelijk dat het aannemelijk is dat er sprake is van unidimensionaliteit en dat deze gekalibreerde opgavenbank één latente trek meet. Dat het bij deze latente trek om de vaardigheid gaat die we ‘begrijpend luisteren’ noemen, blijkt – naast de conclusies met betrekking tot de inhoudsvaliditeit in de vorige paragraaf – uit de resultaten van analyses die we in de rest van dit hoofdstuk presenteren.

6.2.2 Itemkwaliteit

In deze paragraaf vatten we in tabel 6.1 een aantal gegevens samen die betrekking hebben op de itemparameters van de toets Begrijpend luisteren groep 8. Voor een overzicht van alle gegevens per item, zie bijlage 2.
De gemiddelde moeilijkheidsgraad van de toets ligt op het (vooraf) gewenste niveau, namelijk voor M8 tussen de 0,65 en 0,75 (gemiddelde p-waarde is 0,73). De gemiddelde moeilijkheidsgraad voldoet daarmee aan het gestelde doel, namelijk een optimaal onderscheidend vermogen bij de groep met een lage tot gemiddelde vaardigheid (zie verder hoofdstuk 5 over lokale meetnauwkeurigheid), terwijl de toets niet als moeilijk zal 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 toets opgenomen.
De samenhang tussen item- en totaalscore is zowel in termen van R_{ir} als in termen van R_{it} weergegeven. Eerstgenoemde kengetallen geven een reëlere inschatting van die samenhang, maar er zijn geen normwaarden voor beschikbaar in het COTAN-beoordelingssysteem (Evers et al, 2010); voor Rit is dat wel het geval. De gemiddelde Rit-waarde (0,34) is te kenschetsen als ‘goed’. Geen enkele opgave heeft een lagere Rit-waarde dan 0,20.

Tabel 6.1 Samenvatting itemkenmerken voor de toets Begrijpend luisteren M8

	M8		
	p	**R_{it}**	**R_{ir}**
gemiddeld	0,73	0,34	0,25
P10	0,50	0,27	0,19
Mediaan	0,74	0,32	0,22
P90	0,86	0,41	0,33
R_{it} < .20		0	

Zoals vermeld in paragraaf 4.4.2 is voor de toets M8 ook de constante ‘c’ berekend. De nauwkeurigheid van de itemparameterschattingen is voor alle opgaven als ‘goed’ te kenschetsen.

6.2.3 Convergente en discriminante validiteit

Wanneer we de begripsvaliditeit van de toets Begrijpend luisteren M8 evalueren, kunnen we dit doen door na te gaan in hoeverre de toetsscores samenhang vertonen met de scores op andere leervorderingen-toetsen. Als we daarbij op de eerste plaats toetsen kiezen die variëren in de mate van overlap in meetpretentie, krijgen we op deze wijze zicht op de convergente versus discriminante (of divergente) validiteit. Dit is gebeurd door een aantal taaltoetsen te kiezen uit het Cito Volgsysteem primair onderwijs (LOVS) van de tweede generatie (zie paragraaf 6.2.3.1). Als we daarnaast ook een toets afnemen met precies dezelfde meetpretentie (een taak uit de toets Luisteren 3, zie paragraaf 6.2.3.2), kunnen we iets zeggen over de soortgenootvaliditeit.

6.2.3.1 Samenhangen met andere taaltoetsen

Aan de zogeheten ‘volgscholen’ – dit zijn scholen die hebben toegezegd meerdere keren te willen deelnemen aan de proef- en normeringsonderzoeken – die hebben deelgenomen aan het normerings-onderzoek Begrijpend luisteren M8 is via e-mail gevraagd om gegevens beschikbaar te stellen van de afnames van de toetsen voor de vaardigheden Spelling (Cito, 2010), Woordenschat (Cito, 2012), Begrijpend lezen (Cito, 2010) en Technisch lezen (Leestempo; Cito, 2012) via de geautomatiseerde dataretourfunctie van het Computerprogramma LOVS. Al deze toetsen uit de tweede generatie van het LVS zijn door de COTAN (Evers et al, 2010) op alle relevante onderdelen (criteriumvaliditeit is niet van toepassing) met een goed of voldoende beoordeeld.
Cito dataretour is een exporttool die basisscholen in staat stelt om op vrijwillige basis hun LVS-resultaten naar Cito te sturen voor (interne) onderzoeksdoeleinden. Veel basisscholen gaven gehoor aan de oproep (voor aantallen leerlingen zie tabel 6.2; het aantal afnames voor de toets Leestempo is met N = 69 betrekkelijk klein).

Onze verwachting was dat de samenhang tussen het semantische onderdeel Begrijpend luisteren enerzijds en andere semantische onderdelen (Begrijpend lezen en Woordenschat) anderzijds, groter zou zijn dan de samenhang van Begrijpend luisteren met de meer ‘technische’, niet-semantische taalonderdelen zoals Technisch lezen - Leestempo en Spelling (niet-werkwoorden). Vooral tussen Begrijpend luisteren en Woordenschat verwachtten we een hoge correlatie: woordenschat is immers een belangrijke ondersteunende vaardigheid bij Begrijpend luisteren (zie ook hoofdstuk 2). We verwachtten eveneneens een hoge correlatie tussen Begrijpend luisteren en Begrijpend lezen, omdat beide toetsen ‘tekstbegrip’ meten.

In tabel 6.2 worden de (voor attenuatie gecorrigeerde) correlatiecoëfficiënten gerapporteerd tussen de hierboven genoemde toetsen en Begrijpend luisteren op afnamemoment M8.

Tabel 6.2 Correlaties tussen Begrijpend luisteren M8 en verschillende andere LVS-onderdelen*

	Begrijpend luisteren M8	**Aantal leerlingen M8**
Woordenschat	0,76	290
Begrijpend lezen	0,75	324
Spelling niet-werkwoorden	0,41	317
Technisch lezen (Leestempo)	0,10	69

* Deze correlaties zijn gecorrigeerd voor attenuatie

Uit de tabel blijkt dat de verwachtingen bevestigd worden. De correlaties tussen Begrijpend luisteren en de niet-semantisch taalonderdelen zijn laag (voor Spelling niet-werkwoorden 0,41 en voor Technisch lezen - Leestempo 0,10). De correlaties met de semantische taalonderdelen zijn daarentegen hoog (voor Begrijpend lezen 0,75 en voor Woordenschat 0,76, zie de gemarkeerde cellen in de tabel).
Dat de correlatie tussen begrijpend lezen en begrijpend luisteren toeneemt met de jaren, wordt bevestigd in de literatuur. De studie van Gough, Hoover en Peterson (1996) bijvoorbeeld, een studie naar de samenhang tussen technisch lezen en leesbegrip, laat zien dat laat zien dat de samenhang tussen technisch lezen en leesbegrip gelijkmatig afneemt naarmate leerlingen ouder worden. Daarentegen neemt de correlatie tussen begrijpend luisteren en leesbegrip juist toe naarmate er sprake is van een hogere jaargroep: van groep 4 naar groep 6 van 0,60 naar 0,79, daarna blijft de correlatie rond 0,75 in de bovenbouw: de maximale integratie lijkt dan bereikt. Zie hiervoor tabel 6.3 en de wetenschappelijke verantwoordingen Begrijpend luisteren van groep 4 t/m 7 (Van Berkel e.a. 2015 e.v.). Van der Leij (2003) noemt als reden hiervoor dat bij oudere leerlingen begrijpend lezen in grotere mate wordt bepaald door taalbegrip in al zijn aspecten, zoals door woord- en wereldkennis van de leerlingen. Deze bronnen van kennis spelen een belangrijke rol bij het begrijpen van zowel gesproken als geschreven tekst.

Tabel 6.3 Correlaties tussen Begrijpend luisteren en Begrijpend lezen over jaargroepen*

	Groep 4	**Groep 5**	**Groep 6**	**Groep 7**	**Groep 8**
Correlatie Begrijpend luisteren en Begrijpend lezen	0,60	0,72	0,79	0,76	0,75

*Deze correlaties zijn gecorrigeerd voor attenuatie

6.2.3.2 Soortgenootvaliditeit

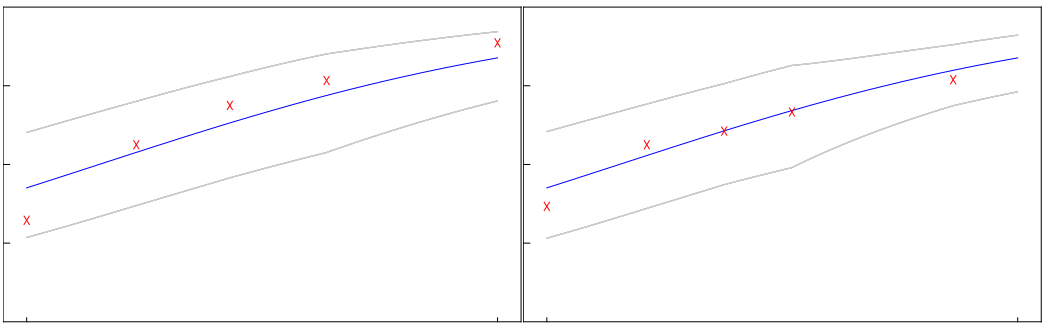
Het is niet eenvoudig om een instrument te vinden dat geschikt is om de soortgenootvaliditeit van de toets Begrijpend luisteren M8 te helpen onderbouwen: er is geen andere toets luistervaardigheid voor het primair onderwijs op de markt die door de COTAN of de Expertgroep toetsen PO als positief is beoordeeld.
Een deel van de leerlingen die hebben meegedaan aan het normeringsonderzoek M8 (172 leerlingen van acht scholen) hebben daarom zowel opgaven van de toets Begrijpend luisteren M8 gemaakt, als een taak (met vijftien opgaven) uit de toets LOVS-Luisteren 3 (Cito, 1996).
De voor attenuatie gecorrigeerde correlatie tussen beide scores is bijzonder hoog: 0,97, zeker als men in aanmerking neemt dat de LVS-toets Begrijpend luisteren kijk- en luisteropgaven bevat en de toets Luisteren 3 uitsluitend luisteropgaven. De toetsen verschillen dus qua vorm van elkaar (er is geen sprake van inhoudelijke overlap; beide toetsen hebben hun eigen opgaven). Dit maakt duidelijk dat er voor deze groep geen onderscheid is tussen audio en video.

6.2.4 Itembias

Er is onderzoek uitgevoerd naar differentieel itemfunctioneren (*Differential Item Functioning,* DIF) met betrekking tot sekse. Voor alle toetsopgaven zijn geobserveerde en verwachte scores voor zowel jongens als meisjes in verschillende scoregroepen berekend. Vervolgens is hier een S-statistiek voor berekend, analoog aan hoe dit gebeurt tijdens de kalibratie (zie hoofdstuk 4).
Het onderzoek naar DIF over afnamemoment en sekse per item liet bij geen enkel item significante verschillen zien op het 1%-niveau. In figuur 6.1 is bij wijze van illustratie van de werkwijze voor een representatief item de toetsing grafisch weergegeven.

Figuur 6.1 Voorbeeld S-toets voor een representatieve opgave uit de M8 toets uitgesplitst naar sekse

M8 jongens **M8 meisjes**



6.2.5 Verschillen tussen relevante subgroepen

Bij de normeringsonderzoeken is het geslacht van de leerlingen opgevraagd. De gemiddelde score van jongens en meisjes is per afnamemoment weergegeven in tabel 6.3.

Tabel 6.3 Verschillen tussen jongens en meisjes voor de toets Begrijpend luisteren M8

M8

	N	M	SD	Effectgrootte *d*
jongen	393	85,15	11,10	0,19
meisje	440	87,29	11,17	

In de tabel is te zien dat meisjes enigszins hoger scoren dan jongens. In termen van effectgrootte is er voor M8 sprake van een klein effect (d = 0,19). Dit is naar verwachting. Op 'talige onderdelen' scoren meisjes in het algemeen iets beter dan jongens (Inspectie van het onderwijs, 2011), maar de verschillen zijn, hoewel consistent, bescheiden. Over de verschillen tussen jongens en meisjes specifiek voor luistervaardigheid is overigens weinig bekend.

Ten slotte besteden we hier – in het kader van de validiteit – nogmaals aandacht aan een belangrijke functie van een leerlingvolgsysteem: het beschrijven en volgen van de vaardigheidsontwikkeling over tijd. In paragraaf 2.3 hebben we hier al aandacht aan besteed door de gemiddelde vaardigheidsscores voor alle afnamemomenten in groep 4 tot en met 8 te presenteren. We veronderstellen dat er tussen de afname-momenten (gemiddeld) sprake is van toename in vaardigheid. Let wel, het gaat hier per leerjaar steeds om verschillende toetsen, die met behulp van IRT op één en dezelfde vaardigheidsschaal zijn gebracht.
We geven de resultaten hier niet nog een keer, maar volstaan met een samenvatting ervan. Na een relatief flinke groei tussen M4 en E4, is er in de leerjaren 5, 6 en 7 steeds sprake van een bescheiden groei, die zich in leerjaar 7 lijkt te stabiliseren. Alleen in groep 8 is er sprake van een grotere groei. Waarom dit zo is, blijft giswerk, maar de eerder gegeven verklaring door Van der Leij (2003) in paragraaf 6.2.3.1 zou plausibel zijn: dat bij oudere leerlingen bij begrijpend lezen het taalbegrip in grotere mate bepaald wordt door woord- en wereldkennis van de leerlingen en dat deze bronnen van kennis spelen een belangrijke rol bij het begrijpen van zowel *gesproken* als geschreven tekst.Bovendien bereiken in de bovenbouw steeds meer leerlingen het punt waarop zij hun intrede doen in de puberteit, waarbij ook hun cognitief functioneren een kwalitatief ander karakter krijgt (vergelijk Piagets formeel-operationele stadium).
Dat de groei van het M-moment naar het E-moment kleiner is dan de groei van het E-moment naar het M-moment, is te verklaren door het feit dat er minder tijd zit tussen het medio- en het eindmoment (4 maanden) dan tussen het eindmoment en het mediomoment van het jaar erop (7 maanden).

Het feit dat de stijgende gemiddelden aansluiten bij de verwachtingen omtrent de vaardigheidsgroei van de leerlingen, interpreteren we als additionele evidentie met betrekking tot de validiteit van de toetsen Begrijpend luisteren, ook die van de toets voor groep 8.

De resultaten geven aan dat de toets Begrijpend luisteren voor groep 8 prima past binnen de reeks van toetsen in het Cito Volgsysteem primair en speciaal onderwijs die bedoeld zijn om de vaardigheid begrijpend luisteren in kaart te brengen en te volgen.

7 Samenvatting

In dit hoofdstuk wordt kort weergegeven wat in de voorafgaande hoofdstukken besproken is.
In hoofdstuk 2 hebben we de uitgangspunten bij de toetsconstructie beschreven en hebben we beschreven hoe de groei van leerlingen met de toetsen Begrijpend luisteren gevolgd kan worden. We zijn hierbij ingegaan op de vraag hoe groei in het Cito Volgsysteem voor primair en speciaal onderwijs wordt gemeten, hoe deze groei aan de hand van de verstrekte gegevens kan worden vastgesteld en geïnterpreteerd, en wat de rol van de toetsbetrouwbaarheid in dit opzicht is. In hoofdstuk 3 hebben we de inhoud van de toets uitvoerig beschreven en verantwoord en hebben we de toets kort gekarakteriseerd in termen van enkele statistische parameters. Vervolgens hebben we in hoofdstuk 4 over de proeftoetsing en het normeringsonderzoek gerapporteerd.
De volgsysteemtoets Begrijpend luisteren voor groep 8 is een toets genormeerd voor één afnamemoment in het schooljaar, het zogeheten M-moment dat halverwege het schooljaar valt. We hebben in hoofdstuk 4 verantwoord hoe het macrodesign en het afnamedesign voor het kalibratieonderzoek is opgezet. Ook hebben we aangegeven hoe we te werk zijn gegaan bij de steekproeftrekking. De wijze van steekproeftrekken en de controles achteraf (wat betreft het percentage achterstandsleerlingen, schoolgrootte, regioverstedelijking en sekse) wijzen uit dat de normeringssteekproef representatief is voor de populatie van scholen in Nederland.
In hoofdstuk 5 rapporteerden we over betrouwbaarheid en meetnauwkeurigheid.
De betrouwbaarheidscoëfficiënt (MAcc's) is voor de toets Begrijpend luisteren als voldoende te interpreteren in het licht van de functie van deze toets. Daarnaast maakt het gehanteerde IRT-model het mogelijk om na te gaan hoe het is gesteld met de lokale meetnauwkeurigheid van de toetsen. Deze is in de lagere en gemiddelde vaardigheidsregionen wat groter dan in de hogere. De betrouwbaarheidstabellen laten zien dat er behoorlijk wat leerlingen zijn die op basis van hun geschatte vaardigheidsscore een niveaugroep te hoog of te laag geplaatst worden, maar de uitkomsten liggen redelijk in lijn met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969). De toets Begrijpend luisteren M8 weet vooral de laagst en hoogst scorende leerlingen accuraat te classificeren; in het midden is de accuraatheid van de classificatie minder. De toch al kleine verschillen in vaardigheidsgroei tussen de afnamemomenten moeten tegen de achtergrond van de meetfouten worden geïnterpreteerd en uitspraken over progressie in termen van vaardigheidsniveaus dienen met voorzichtigheid te worden gehanteerd.
Over validiteit rapporteerden we in hoofdstuk 6. Voor toetsen in een leerlingvolgsysteem is de inhoudsvaliditeit van de toetsen van buitengewoon groot belang. De basis is daarvoor gelegd in hoofdstuk 2 en 3. Uitgangspunten bij de toetsconstructie waren de kerndoelen primair onderwijs en de tussendoelen Mondelinge communicatie en leerlijnen van TULE. Dit heeft geresulteerd in een indeling in de vaardigheden 'Begrijpen' en 'Interpreteren', vaardigheden die in de werkelijkheid niet altijd duidelijk van elkaar te scheiden zijn. Daarnaast zijn verschillende opgaventypen onderscheiden die bij de toetsconstructie leidend waren. Het referentiekader Taal was bij de ontwikkeling van de uitgangspunten van de toetsreeks nog niet verschenen Direct na verschijnen is nagegaan hoe de uitgangspunten van de toetsen Begrijpend luisteren zich verhouden tot wat in het referentiekader beschreven werd bij 'Luisteren'. De uitgangspunten bleken in grote lijnen overeen te komen. De constructie van teksten en opgaven was in handen van ervaren leerkrachten basisonderwijs in samenwerking met toetsdeskundigen van Cito.
Door de toetsen evenwichtig en in overeenstemming met de uitkomsten van de analyses samen te stellen, door een adequate itemconstructiegroep en adequate itemconstructieprocedures in te zetten en door het uitvoeren van proeftoetsingen kunnen we uiteindelijk concluderen dat er sprake is van een inhoudsvalide toets die aansluit bij het niveau begrijpend luisteren van leerlingen in groep 8.
Daarnaast is uitgebreid ingegaan op de begripsvaliditeit van de toets Begrijpend luisteren groep 8.
Een belangrijke indicatie voor de validiteit van de opgaven uit de toets komt uit het kalibratieonderzoek (hoofdstuk 4). Daaruit is gebleken dat de opgavenverzameling waaruit de toets is samengesteld, beschreven kan worden met OPLM. Dat betekent dat de met de toets gemeten vaardigheid te verklaren is door een unidimensionaal model. In concreto betekent dit dat alle toetsopgaven een beroep doen op dezelfde (veronderstelde, latente) vaardigheid.

Een belangrijke aanwijzing voor de convergente en discriminerende validiteit is af te leiden uit de correlaties tussen de toets Begrijpend luisteren met andere toetsen uit het Cito Volgsysteem primair en speciaal onderwijs. Uit deze gegevens blijkt dat de scores op de toets Begrijpend luisteren sterk samenhangen met scores op semantische taalonderdelen, zoals woordenschat en begrijpend lezen, en nauwelijks met scores op de andere, niet-semantische taalonderdelen, zoals technisch lezen en spelling. Een andere belangrijke aanwijzing voor begripsvaliditeit betreft de zeer hoge correlatie met een soortgenoot: een taak uit de toets LOVS-Luisteren 3.
De gegevens over de itemkenmerken (moeilijkheidsgraad en item-totaalcorrelatie) laten zien dat de itemkwaliteit bevredigend is: alle items voldoen aan de daarvoor geldende kwaliteitscriteria. De gemiddelde p-waarde voor de toets is 0,73. De gemiddelde R_{it}-waarden zijn te kenschetsen als ‘goed’ (gemiddelde $R_{it} > 0,30$). Dit laatste valt ook af te leiden uit de conclusies van de analyses die zijn uitgevoerd met betrekking tot de schatting van de nauwkeurigheid van de itemparameters (op basis van constante ‘c’).
Uit het onderzoek dat is uitgevoerd naar differentieel itemfunctioneren blijkt dat er voor sekse en afnamemoment geen sprake is van itembias.
De gemiddelde scores op Begrijpend luisteren M8 naar sekse laten zien dat meisjes als groep iets hoger scoren. Deze bevinding sluit aan bij het gegeven dat meisjes op talige toetsonderdelen over het algemeen enigszins in het voordeel zijn ten opzichte van jongens. De verschillen zijn echter klein. Ten slotte konden we laten zien dat de toets Begrijpend luisteren M8 goed past in de reeks vergelijkbare toetsen in de groepen 4 tot en met 8 die bedoeld zijn om deze vaardigheid te beschrijven en te volgen. Er is steeds sprake van een lichte vaardigheidsgroei, die zich aanvankelijk leek te stabiliseren in de hogere groepen (groep 6 en 7). Echter, voor groep 8 is de vaardigheidsgroei wat groter dan verwacht. Mogelijk wordt in groep 8 bij begrijpend luisteren het taalbegrip in grotere mate bepaald door woord- en wereldkennis van de leerlingen en veranderingen in het cognitief functioneren die zij doormaken.

8 Literatuur

Aarnoutse, C. & L. Verhoeven (2003). *Tussendoelen gevorderde geletterdheid. Leerlijnen voor groep 4 tot en met 8.* Nijmegen: Expertisecentrum Nederlands.

Bachman, L. (1990). *Fundamental considerations in language testing*. Chapter 5 (pp. 111-159). Oxford: Oxford University Press.

Berkel, S. van, F. van der Schoot, R. Engelen en G. Maris (2002). *Balans van het taalonderwijs halverwege de basisschool 3. Uitkomsten van de derde taalpeiling in 1999.* Arnhem: Cito (PPON-reeks nr. 20).

Berkel, S. van, I. Groenen en M. Hilte (2013). *Woordenschat Groep 8.* Leerling- en onderwijsvolgsysteem primair onderwijs. Arnhem: Cito.

Berkel, S. van, R. Engelen, M. van Groen, M. Hilte, M. van der Zanden, (2013). *Wetenschappelijke verantwoording Begrijpend luisteren groep 3. Cito Volgsysteem primair en speciaal onderwijs.* Arnhem: Cito.

Berkel, S. van, R. Engelen, M. Hilte, J. Wouda, M. van der Zanden, (2015). *Wetenschappelijke verantwoording Begrijpend luisteren groep 4. Cito Volgsysteem primair en speciaal onderwijs.* Arnhem: Cito.

Berkel, S. van, R. Engelen, M. Hilte, J. Wouda, M. van der Zanden, (2015). *Wetenschappelijke verantwoording Begrijpend luisteren groep 5. Cito Volgsysteem primair en speciaal onderwijs.* Arnhem: Cito.

Berkel, S. van, R. Engelen, M. Hilte, F. Kamphuis, M. van der Zanden, (2015). *Wetenschappelijke verantwoording Begrijpend luisteren groep 6. Cito Volgsysteem primair en speciaal onderwijs.* Arnhem: Cito.

Berkel, S. van, R. Engelen, M. Hilte, F. Kamphuis, M. van der Zanden, (2016). *Wetenschappelijke verantwoording Begrijpend luisteren groep 7. Cito Volgsysteem primair en speciaal onderwijs.* Arnhem: Cito.

Besluit 551 (2005). Besluit vernieuwde kerndoelen WPO. *Staatsblad van het Koninkrijk der Nederlanden.*

Besluit 283 (2006). Besluit van 19 mei 2006, houdende wijziging van het Besluit bekostiging WPO in verband met een wijziging van de gewichtenregeling en wijziging van het Besluit bekostiging WEC in verband met een wijziging in de groepsgrootte. *Staatsblad van het Koninkrijk der Nederlanden.*

Bostrom, R.N. (1990). *Listening behavior. Measurement and application.* New York/London: The Guilford Press.

Bostrom, R.N. (1997). The testing of mother tongue listening skills. In: Clapham, C. and D. Corson (Eds.) *Encyclopedia of language and education. Volume 7. Language testing and assessment* (pp. 21-27). Dordrecht: Kluwer.

Buck, G. (1989). *Listening comprehension: construct validity and trait characteristics.* Paper 11th Language testing Research Colloquium: San Antonio.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing*, 8, 67-91.

Buck, G. (2001). *Assessing listening.* Cambridge, UK: University Press.

Chang, A.C. en J. Read (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly*, 40, 375-397.

Cito (2011). *Begrijpend luisteren groep 3*. Cito Volgsysteem primair onderwijs (LVS). Arnhem: Cito.

Cito (2012). *Begrijpend luisteren groep 4*. Cito Volgsysteem primair onderwijs (LVS). Arnhem: Cito.

Cito (2013). *Begrijpend luisteren groep 5*. Cito Volgsysteem primair onderwijs (LVS). Arnhem: Cito.

Cito (2014). *Begrijpend luisteren groep 6*. Cito Volgsysteem primair onderwijs (LVS). Arnhem: Cito.

Cito (2015). *Begrijpend luisteren groep 7*. Cito Volgsysteem primair onderwijs (LVS). Arnhem: Cito.

Damhuis, R. & P. Litjens (2003): *Mondelinge Communicatie, drie werkwijzen voor mondelinge taalontwikkeling.* Nijmegen: Expertisecentrum Nederlands.

Eggen, T.J.H.M., (1993). Itemresponstheorie en onvolledige gegevens. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 239-284). Arnhem: Cito.

Engelen, R.J.H. en Eggen, T.J.H.M. (1993). Equivaleren. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk.* (pp. 239-284). Arnhem: Cito.

Evers, A., W. Lucassen, R. Meijer en K. Sijtsma (2010), *COTAN Beoordelingssysteem voor de kwaliteit van tests*. Amsterdam: NIP/COTAN.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008a). *Over de drempels met taal en rekenen. Hoofdrapport van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen*. Enschede: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008b). *Over de drempels met taal. De niveaus voor de taalvaardigheid*. Enschede: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009a). *Referentiekader taal en rekenen. De referentieniveaus.* Enschede: Expertgroep doorlopende leerlijnen Taal en Rekenen.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2009b). *Een nadere beschouwing. Over de drempels met taal en rekenen*. Enschede: Expertgroep doorlopende leerlijnen Taal en Rekenen.

Feenstra, H. (2010). *Begrijpend lezen Groep 8*. Leerling- en onderwijsvolgsysteem primair onderwijs. Arnhem: Cito.

Friedman, S.J. en T.N. Asley (1990). The influence of reading on listening test scores. *Journal of Experimental Education*, 58, 301-310.

Gijsel, M. en M. van Druenen (2011), *Opbrengstgericht werken aan mondelinge taalvaardigheid*. Nijmegen: Expertisecentrum Nederlands.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, [Electronische versie], 19, 133-166.

Glas, C.A.W. & Verhelst, N.D., (1993). Een overzicht van itemresponsmodellen. In: T.J.H.M. Eggen & P.F. Sanders (red.). Psychometrie in de praktijk. (pp. 179-238). Arnhem: Cito.
Gough P. B., Hoover W. A., Peterson C. (1996). Some observations on the simple view of reading. In: Cornoldi C., Oakhill J. (Eds.), Reading comprehension difficulties. Hillsdale, NJ: Erlbaum.

Greven, J., J. Letschert, SLO (2006). *Kerndoelen primair onderwijs*, Publicatie van het ministerie van Onderwijs, Cultuur en Wetenschap.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item response Theory*. Newbury Park, CA: Sage.

Hemker, B.T., J. Kordes & J.J. van Weerden (2011). *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2010 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau*. Arnhem: Cito.

Heuvelman, A. en K. Schreuder (1994). Luisteren met open ogen. Factoren in de verwerking van audiovisuele informatie. *Tijdschrift voor Taalbeheersing*, 16, 32-45.

Hollenberg, J. en J. Vloedgraven, (2012). *Wegwijzer toetsgebruik bij leerlingen met extra onderwijsbehoeften/speciale leerlingen*. Arnhem: Cito.

Inspectie van het Onderwijs, Ministerie van OCW, *De staat van het onderwijs*, Onderwijsverslag 2009/2010. Utrecht: 2011

Jongen, I., R. Krom en P. Roumans (2012), *Technisch lezen Groep 8, Leestechniek en Leestempo.* Leerling- en onderwijsvolgsysteem primair onderwijs, Arnhem: Cito.

Krom, R.S.H. (1996). *Luisteren 3*. Arnhem: Cito.

Krom, R. (1997). Het verbeteren van de luisterhouding in de klas. In: *Gids voor het Basisonderwijs*, 40e aanvulling. Diegem: Kluwer Editorial (Wolters Kluwer NV).

Krom, R.S.H., Ouborg, M.J., & Kamphuis, F.H. (2001). *Wetenschappelijke verantwoording van de toetsseries Luisteren 1, 2 en 3. Leerlingvolgsysteem*. Arnhem: Citogroep.

Krom, R.S.H., S. van Berkel, F. van der Schoot, J. Sijtstra, B. Hemker en M. Marsman (2011). *Balans van het luisteronderwijs in het basis- en speciaal basisonderwijs. Uitkomsten van de vierde peiling in 2007.* Arnhem: Cito (PPON-reeks nr. 46).

Leij, A. van der (2003) *Leesproblemen en dyslexie: Beschrijving, verklaring en aanpak.* Rotterdam: Lemniscaat.

Levelt, W.J.M. (1989). *Speaking. From Intention to Articulation*. Cambridge, Mass: MIT.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Ministerie van Onderwijs, Cultuur en Wetenschappen (2006). *Kerndoelenboekje*. www.minocw.nl

Nulft, D. van den & Verhallen, M. (2002). *Met woorden in de weer. Woordenschatuitbreiding en cognitieve ontwikkeling van leerlingen.* Bussum: Coutinho.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Rost, M. (1999). Listening in a Second Language. In Spolsky, B. (Ed.), *Concise Encyclopedia of Educational Linguistics* (pp. 290-295). Amsterdam: Elsevier.

Osada, N. (2004). Listening comprehension research: a brief review of the past thirty years. *Dialogue*, 3, 53-66.

Poelmans, P. (2003). *Developing second-language listening comprehension: effects of training lower-order skills versus higher-order strategy*. Dissertatie Universiteit van Amsterdam.

Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes*, [Electronische versie], 1, 105-119.

Richards, J.C. (1983). Listening comprehension: approach, design, procedure. *TESOL Quarterly*, 17, 219-240.

Samuels, S.J. (1987). Factors that influence listening and reading comprehension. In: R. Horowitz and S.J. Samuels (Eds.), *Comprehending oral and written language*. San Diego, etc.: Academic Press.

Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing*, 14, 185-213.

Shohamy, E. en O. Inbar (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing*, 8, 23-40.

Sijtstra, J., F. van der Schoot en B. Hemker (2002). *Balans van het taalonderwijs aan het einde van de basisschool 3. Uitkomsten van de derde peiling in 1998*. Arnhem: Cito (PPON-reeks nr. 19).

Sijtstra, J. (2005). *Domeinbeschrijving luistervaardigheid*. Intern stuk, Arnhem: Cito

Spearitt, D. (1999). Language Testing in Mother Tongue. In Spolsky, B. (Ed.), *Concise Encyclopedia of Educational Linguistics* (pp. 715-721). Amsterdam: Elsevier.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid: de ontwikkeling van een domeingericht meetinstrument.* Enschede: Universiteit Twente.

Staphorsius, G., Krom, R.S.H., Kleintjes, F.G.M & N.D. Verhelst (2004). *Verantwoording van de Toetsen Begrijpend Lezen (TBL)*. Arnhem: Citogroep.

Tannen, D. (1982). Spoken and written language. Exploring orality and literacy. New Jersey, Ablex.
Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, 24, 3-25.

Tomesen, M. en De Wijs, A. (2010), *Spelling Groep 8.* Leerling- en onderwijsvolgsysteem primair onderwijs, Arnhem: Cito.

TULE, Tomesen, M.A. (2008). *TULE - Nederlands: Inhouden en activiteiten bij de kerndoelen van 2006*. Enschede: SLO.

Verhallen, M. & Verhallen, S. (1994). *Woorden leren woorden onderwijzen.* Hoevelaken: CPS.

Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.H.J.M. (1991). Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model. *Measurement and Research Department Reports 91-10*. Arnhem: Cito.

Verhelst, N.D. (1992). *Het één parameter model (OPLM). Een theoretische inleiding en een handleiding bij het computerprogramma*. Arnhem: Cito.

Verhelst, N.D. (1993). Itemresponstheorie. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 83-178). Arnhem: Cito.

Verhelst, N.D. & Kleintjes, F.G.M. (1993). Toepassingen van itemresponsetheorie. In: T.J.H.M. Eggen en P.F. Sanders (Red.). *Psychometrie in de praktijk*. Arnhem: Cito.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In: G.H. Fischer & I.W. Molenaar (Eds.). *Rasch models: Foundations, recent developments and applications* (pp. 215-239). New York: Springer.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1995). *OPLM: One Parameter Logistic Model. Computer program and manual*. Arnhem: Cito.

Verhoeven, L., H. Biemond en P. Litjens (2007). *Tussendoelen mondelingen communicatie. Leerlijnen voor groep 1 tot en met 8. Nijmegen*, Expertisecentrum Nederlands.

Verstralen, H.H.F.M. (1997). *OPTAL: Inverse OPLAT and item and test characteristics in populations.* Arnhem, The Netherlands: Cito.

Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs. Een inventarisatie van beoordelingsmethoden voor de stelvaardigheid, het begrijpend lezen, de spreek-, luister- en discussievaardigheid*. Den Haag: SVO.

Widdowson, H.G. (1990). Aspects of language teaching. Oxford, Oxford University Press. Wilson, M. (2003). Discovery listening – improving perceptual processing. *ELT Journal,* 57, 335-343.

Yi'an, W. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, [Electronische versie], 15, 21-44.

Bijlagen

Bijlage 1 Kerndoelen Nederlands PO

Mondeling taalonderwijs
1. De leerlingen leren informatie te verwerven uit gesproken taal. Ze leren tevens die informatie, mondeling of schriftelijk, gestructureerd weer te geven.
Taalbeschouwing
12. De leerlingen verwerven een adequate woordenschat en strategieën voor het begrijpen van voor hen onbekende woorden. Onder 'woordenschat' vallen ook begrippen die het leerlingen mogelijk maken over taal te denken en te spreken.

Bijlage 2 Items en waarden toets M8

Marg			M8	
InBk	Label	P-Val	RIT	RIR
1	1301	0,68	0,31	0,21
2	1295	0,52	0,32	0,22
3	1125	0,68	0,31	0,21
4	1128	0,80	0,35	0,27
5	1285	0,87	0,23	0,16
6	993	0,82	0,34	0,26
7	994	0,61	0,47	0,38
8	1286	0,84	0,33	0,25
9	1084	0,73	0,30	0,21
10	1086	0,80	0,35	0,27
11	1087	0,67	0,31	0,21
12	1091	0,82	0,26	0,18
13	1293	0,90	0,34	0,28
14	1291	0,46	0,32	0,22
15	1294	0,74	0,37	0,29
16	1029	0,60	0,32	0,22
17	1030	0,83	0,33	0,26
18	1032	0,49	0,40	0,31
19	1033	0,43	0,32	0,22
20	1275	0,89	0,29	0,22
21	1058	0,79	0,27	0,19
22	1059	0,81	0,26	0,18
23	1060	0,80	0,41	0,34
24	1061	0,83	0,45	0,38
25	1276	0,84	0,39	0,32
26	1063	0,76	0,37	0,28
27	1278	0,73	0,29	0,20
28	891	0,76	0,37	0,28
29	1254	0,73	0,38	0,29
30	1256	0,73	0,29	0,20
31	1148	0,62	0,32	0,22
32	896	0,64	0,31	0,22
33	1257	0,74	0,37	0,29

Cito helpt je inzicht te krijgen in je ontwikkeling en mogelijkheden. Door kennis, vaardigheden en competenties objectief meetbaar te maken en de ontwikkeling er van te volgen, kun je het beste uit jezelf halen, verantwoorde keuzes maken en beter richting geven aan je toekomst. Cito draagt daaraan bij door wereldwijd werk te maken van goed en eerlijk toetsen, vanuit de kernwaarden kundig, toonaangevend, integer, innovatief en betrokken.

Cito
Amsterdamseweg 13
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
www.cito.nl

Fotografie: Ron Steemers