Wetenschappelijke verantwoording Begrijpend luisteren groep 3

Saskia van Berkel, Ronald Engelen, Maaike van Groen, Maartje Hilte, Jasper Wouda en Mart van der Zanden

zeker weten

Wetenschappelijke verantwoording

Begrijpend luisteren groep 3

Saskia van Berkel
Ronald Engelen
Maaike van Groen
Maartje Hilte
Jasper Wouda
Mart van der Zanden

Cito
Arnhem, 2013

Inhoud

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de toetsen Begrijpend luisteren voor groep 3. De toetsen Begrijpend luisteren maken deel uit van de tweede generatie toetsen van het Cito Volgsysteem primair onderwijs en zijn primair bestemd voor leerlingen in de groepen 3 t/m 8 in het primair onderwijs. Het betreft papieren toetsen[1] voor alle leerjaren.
Te zijner tijd zullen ook de wetenschappelijke verantwoordingen met de gegevens van de (nog te verschijnen) toetsen Begrijpend luisteren voor de groepen 4 t/m 8 gefaseerd worden uitgebracht.

Deze verantwoording biedt tezamen met de inhoud van het toetspakket Begrijpend luisteren voor groep 3 alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit van de betreffende meetinstrumenten. Het genoemde materiaal maakt een beoordeling van de toetsen Begrijpend luisteren 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 onderwijs 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 begripsvaliditeit (hoofdstuk 6) van de toetsen Begrijpend luisteren voor groep 3. 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 onderwijs zullen er geen digitale toetsen voor Begrijpend luisteren worden 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; Gijssel & Van Druenen, 2011). De toetsen Begrijpend luisteren voor groep 3 van het Cito Volgsysteem primair onderwijs beogen de vaardigheid in begrijpend luisteren te meten en de opgaven in de toetsen zijn een operationalisering van deze vaardigheid. De toetsen bevatten geen opgaven in de sfeer van de traditionele auditieve discriminatietaken waarin leerlingen moeten luisteren met als doel het onderscheiden van geluiden, klanken of stemmen.
De toetsen Begrijpend luisteren zijn bedoeld om vast te stellen hoe de vaardigheid begrijpend luisteren van leerlingen zich ontwikkelt van groep 3 tot en met groep 8 en maken inzichtelijk welke verschillen in luistervaardigheid er tussen leerlingen bestaan. Met behulp van de toetsen kan onderscheid gemaakt worden tussen vaardige en minder vaardige luisteraars en kunnen de vorderingen van de leerlingen met betrekking tot hun luistervaardigheid in kaart worden gebracht (zie verder paragraaf 2.4.1).

2.2 Doelgroep

De toetsen Begrijpend luisteren voor groep 3 zijn primair bestemd voor en genormeerd bij leerlingen in groep 3 van het Nederlandse basisonderwijs. De populatieparameters voor de toetsen zijn zowel op het midden als op het einde van het schooljaar bepaald. Desgewenst kunnen de toetsen 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 toetsen zijn ook geschikt voor leerlingen op speciale scholen voor basisonderwijs en voor speciale leerlingen in het regulier basisonderwijs. In de handleiding bij de toetsen zijn – met het oog op het gebruik voor deze leerlingen – extra aanwijzingen opgenomen. Er zijn echter geen aparte referentiegegevens verzameld: voor speciale leerlingen zijn dezelfde normen van toepassing als voor reguliere leerlingen. Daardoor zijn de prestaties van beide doelgroepen op de toetsen (speciaal versus regulier) vergelijkbaar.
Voor leerlingen die nog maar pas in Nederland verblijven, zijn de toetsen ongeschikt: leerlingen dienen minstens vier jaar deel te hebben genomen aan de Nederlandse samenleving (kinderopvang/peuter-speelzaal//basisonderwijs), voordat de toetsen Begrijpend luisteren bij hen mogen worden afgenomen.
De toetsen zijn ook niet geschikt voor leerlingen met gehoorproblemen.

2.3 Gebruiksdoel en functie

De toetsen Begrijpend luisteren in het Cito Volgsysteem primair onderwijs 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 die van een grote groep leerlingen.
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
De toetsen Begrijpend luisteren van het Cito Volgsysteem geven de leerkracht informatie over de ontwikkeling van de luistervaardigheid van de leerlingen, individueel en als groep, gedurende (vrijwel) de gehele basisschoolperiode. 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?
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 – 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).

2.4 Theoretische inkadering

2.4.1 Theoretische inkadering: inhoudelijk

Wat is 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. Het gaat dan nadrukkelijk niet om het 'technisch' luisteren, waarbij leerlingen in de onderbouw van het basisonderwijs aan de hand van auditieve discriminatietaken geluiden, klanken of stemmen moeten onderscheiden.

Karakteristieken van gesproken taal
Gesproken taal kent een aantal belangrijke karakteristieken (Buck, 2001). Het 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. Het 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 ideeëneenheden met een eenvoudige grammaticale structuur, die bijeengehouden worden door nevenschikkende verbanden (en, maar, of). Er zijn aarzelingen en pauzes, opvullers en

[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, waarbij het vooral gaat om transactioneel taalgebruik (zie p. 9).

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

Begrijpen, Interpreteren en Reflecteren
Bij luisteren is sprake van interactie tussen de 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[3]. 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, 2009).
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 is een continue wisselwerking. Wanneer tekst- en kennisgestuurde verwerking in samenhang en gelijktijdig ingezet worden, is er pas sprake van werkelijk en diepgaand tekstbegrip. Krom e.a. (2011) 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, oftewel 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*. 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.

[3] De term ´tekst´ is een verzamelterm: hieronder vallen ook de gesproken taaluitingen waarmee de leerlingen in aanraking komen bij het maken van de toetsen Begrijpend luisteren groep 3.

Luistercontext: Interactioneel en transactioneel taalgebruik

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, 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. In de schoolse context vindt vooral (maar niet alleen) transactioneel taalgebruik plaats, waarbij veelal sprake is van eenrichtingsverkeer. 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.

Begrijpend luisteren en andere vaardigheden

Luistervaardigheid is van belang om zowel in de thuisomgeving, in de samenleving als op school 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 mede 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 lezen als bij begrijpend luisteren 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 de lezer als de luisteraar moet de tekst decoderen, begrijpen en interpreteren. Ook het toepassen van linguïstische kennis en het inzetten van achtergrondkennis is zowel bij begrijpend lezen als bij begrijpend luisteren 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, blindvaren op de automatische piloot (cf. Buck, 2001; p. 6).

Uiteraard spelen de specifieke tekst en de context hierbij 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: 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.

2.4.2 Theoretische inkadering: psychometrisch

Opgavenbanken
Voor het samenstellen van toetsen voor het basisonderwijs beschikt Cito over opgavenbanken. Deze liggen onder meer ten grondslag aan de toetsen in het Cito Volgsysteem primair onderwijs, waaronder de LVS-toetsen, de Entreetoetsen en de Eindtoets Basisonderwijs. Voor de constructie van de toetsen Begrijpend luisteren hebben we gebruik gemaakt van de opgavenbank Begrijpend luisteren. Ook voor andere vakgebieden in het Cito Volgsysteem zoals Begrijpend lezen, Woordenschat, Spelling, Rekenen-Wiskunde en Studievaardigheden zijn opgavenbanken in gebruik.

Een opgavenbank is nadrukkelijk niet 'zomaar' een verzameling opgaven waaruit een toetsconstructeur min of meer naar willekeur een aantal opgaven selecteert om een nieuwe toets samen te stellen. We geven hier kort aan wat de vereisten zijn om van een deugdelijke en psychometrisch goed gefundeerde opgavenbank te kunnen spreken.

Unidimensionaal continuüm
Het algemene uitgangspunt is dat de luistervaardigheid 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 uit van de luistervaardigheid, waarbij een groter getal wijst op een grotere vaardigheid. Het doel van de meetprocedure – het afnemen van een 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 is de schatting van de plaats van de leerling op het vaardigheidscontinuüm. De tweede geeft aan hoe nauwkeurig die schatting is en heeft dus de status van een standaardfout, te vergelijken met de standaardmeetfout uit de klassieke testtheorie.

Latente vaardigheid
De antwoorden die een leerling geeft, worden beschouwd als indicatoren van de luistervaardigheid, hetgeen ruwweg betekent dat men verwacht dat alle opgaven in de bank de luistervaardigheid meten.
De vaardigheid zelf wordt als niet observeerbaar beschouwd en daarom gewoonlijk omschreven als een latente vaardigheid.

'Moeilijkheid' in de Item Response Theorie

Hoewel opgaven dezelfde vaardigheid meten, kunnen ze toch systematisch van elkaar verschillen. Het belangrijkste verschil tussen de opgaven is hun moeilijkheidsgraad. In de klassieke testtheorie wordt moeilijkheidsgraad uitgedrukt met een zogenaamde p-waarde, de proportie correcte antwoorden op een opgave in een welbepaalde populatie van leerlingen. In de Item Response Theorie (IRT) die voor het construeren van de opgavenbanken werd gebruikt, hanteert men echter een andere definitie van moeilijkheid: ruwweg gesproken is het de mate van vaardigheid die nodig is om de opgave goed te kunnen beantwoorden. Dit verschil in definitie van de moeilijkheidsgraad tussen de klassieke theorie en IRT is uitermate belangrijk. Men kan verwachten dat de p-waarde van een opgave in groep 4 groter zal zijn dan in groep 3, waardoor duidelijk wordt dat de p-waarde een relatief begrip is: ze geeft de moeilijkheid aan van een opgave in een bepaalde populatie. Binnen de IRT is de moeilijkheid van een opgave gedefinieerd in termen van de onderliggende vaardigheid, zonder enige referentie 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 die in de vorige alinea werd gehanteerd (de mate van vaardigheid nodig om een opgave goed te kunnen beantwoorden) behoeft enige verdere uitwerking. Men zou deze omschrijving kunnen opvatten als een soort drempel: heeft een leerling die mate van vaardigheid niet, dan kan hij de opgave niet juist beantwoorden; heeft hij die drempel wel gehaald, dan geeft hij (gegarandeerd) het juiste antwoord. Deze interpretatie weerspiegelt een deterministische kijk op het antwoordgedrag van de leerling, die echter in de praktijk geen stand houdt, omdat er uit volgt dat een leerling die een moeilijke opgave correct beantwoordt geen fout kan maken op een gemakkelijke opgave. Daarom wordt in de IRT een kansmodel gebruikt: hoe groter de vaardigheid, des te groter de kans dat een opgave juist wordt beantwoord. De moeilijkheidsgraad van een opgave wordt dan gedefinieerd als de mate van vaardigheid die nodig is om met een kans van precies een half een juist antwoord te kunnen geven.

Kalibratie

In het voorgaande zijn nogal wat veronderstellingen ingevoerd (unidimensionaliteit; alle opgaven zijn indicatoren voor dezelfde vaardigheid; kansmodel) die niet zonder meer voor waar kunnen worden aangenomen; we zullen methoden moeten bedenken om aan te tonen dat al die veronderstellingen deugdelijk zijn. Dit 'aantonen' gebeurt met statistische gereedschappen waarop we in deze paragraaf dieper zullen ingaan. Maar voor we de opgaven in een toets kunnen gebruiken moeten we ook proberen 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 die hiervoor wordt gebruikt noemen we kalibratiesteekproef.

Afnamedesigns

Meestal bevat een opgavenbank meer opgaven dan een doorsnee toets. Bij het uittesten van opgaven is het praktisch niet haalbaar, maar ook niet wenselijk om alle opgaven aan alle leerlingen voor te leggen. Elke leerling in de kalibratiesteekproef krijgt daarom slechts een gedeelte van de opgaven uit de opgavenbank voorgelegd. Dit gedeeltelijk voorleggen gebeurt aan de hand van een 'onvolledig design' en moet met de nodige omzichtigheid gebeuren. Verderop wordt ingegaan op het afnamedesign dat voor de kalibratie is gebruikt, de geïnteresseerde lezer wordt verwezen naar Eggen (1993).

Belangrijke implicaties gekalibreerde opgavenverzameling

Als we erin slagen de kalibratie met succes uit te voeren houden we een zogenaamde gekalibreerde opgavenbank over. In dat proces worden de opgaven die niet passen bij de verzameling verwijderd. De opgavenbank bevat voor elke opgave niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen en de statistische zekerheid dat alle opgaven dezelfde vaardigheid aanspreken.

Dit houdt onder meer het volgende in:

1 In principe kunnen we met een willekeurige selectie opgaven uit de opgavenbank de vaardigheid meten bij een willekeurige leerling. In principe, want een willekeurige toets die uit de opgavenbank wordt getrokken zal in de praktijk meestal niet voldoen omdat het meetresultaat (de schatting van de vaardigheid) onvoldoende nauwkeurig zal zijn. Willen we een nauwkeuriger meting (bij een gegeven aantal opgaven in de toets) dan zullen we de moeilijkheidsgraden van de opgaven in overeenstemming moeten brengen met het vaardigheidsniveau van de leerlingen.
2 We kunnen een schatting maken van de verdeling van de vaardigheid in een welomschreven populatie, door selecties van opgaven voor te leggen aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. In het geval van het Cito Volgsysteem primair onderwijs zijn dat steekproeven van leerlingen op de verschillende normeringsmomenten vanaf medio groep 3 (M3) tot en met medio groep 8 (M8). Daarbij maakt het, behoudens wat bij 1 is vermeld over nauwkeurigheid, niet uit welke selectie van opgaven aan een leerling binnen een normeringsgroep wordt afgenomen. Een van de eigenschappen van gekalibreerde opgavenbanken is immers dat met elke opgavenselectie de vaardigheid van leerlingen kan worden bepaald. Zie voor een voorbeeld hiervan Staphorsius (1994). In de praktijk komt dit meestal neer op het schatten van gemiddelde en standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met deze schattingen kunnen dan ook schattingen gemaakt worden van de percentielen in de populatie.
3 Aan leerlingen die niet behoren tot de betreffende referentiepopulatie kan dezelfde toets worden voorgelegd. De toetsscore wordt omgezet in een schatting van de vaardigheid en deze schatting kan geplaatst worden in de vaardigheidsverdeling van de populatie. Een leerling met achterstand in groep 4 kan een toets maken die normaliter aan groep 3 wordt voorgelegd, en zijn of haar vaardigheids-schatting kan behalve met de populatie van groep 4 ook vergeleken worden met de percentielen in de populatie van groep 3, met bijvoorbeeld de uitspraak: "De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 3."
4 De vergelijking die in het voorgaande gemaakt is, kan evengoed plaatsvinden als de (achterstands)-leerling een andere toets maakt (i.e. een selectie uit de opgavenbank) dan de toets die normaliter aan de leerlingen in groep 4 wordt voorgelegd. Immers het kalibratieonderzoek heeft ons overtuigd dat alle opgaven dezelfde vaardigheid meten. Met een nieuwe toets meten we dus dezelfde vaardigheid, zodat schattingen die van verschillende toetsen afkomstig zijn zinvol met elkaar kunnen worden vergeleken.

Tot zover onze nadere bepaling van het begrip 'opgavenbank'. In de volgende hoofdstukken van deze verantwoording worden de begrippen die hierboven aan de orde zijn geweest nader uitgewerkt en toegelicht voor de opgavenbank Begrijpend luisteren. Voor de verantwoording van de constructie van deze opgavenbank verwijzen we naar hoofdstuk 3. In hoofdstuk 6 wordt de validering van de opgavenbanken besproken.

Het gehanteerde meetmodel

In het normeringsonderzoek is gebruikgemaakt van een op de Item Respons Theorie gebaseerd meetmodel zoals dat bij Cito gebruikelijk is. Dergelijke modellen verschillen in een aantal opzichten vrij sterk van de klassieke testtheorie (Verhelst, 1993; Verhelst en Glas, 1995). Bij de klassieke testtheorie staan de toets en de toetsscore centraal. Het theoretisch belangrijkste begrip in deze theorie is de zogenaamde ware score, de gemiddelde score die de persoon zou behalen indien de test een oneindig aantal keren onder dezelfde condities zou worden afgenomen. Die notie geeft een van de belangrijkste (praktische) obstakels van deze theorie voor ons onderzoek weer: het is problematisch om toetsscores te vergelijken die verkregen zijn in een onvolledig design. Hoewel er methoden bestaan binnen de klassieke testtheorie om toetsscores te equivaleren (Engelen & Eggen, 1993), schiet deze benadering tekort als het gaat om de centrale vraag: hoe weten we dat de equivalering zinvol is? Op die vraag heeft IRT een antwoord.

In de IRT staat het te meten begrip of de te meten eigenschap centraal. De IRT beschouwt het antwoord op een opgave als een indicator voor de mate waarin die eigenschap aanwezig is. Het verband tussen eigenschap en antwoord op een opgave is van probabilistische aard en wordt weergegeven in de zogenaamde 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 kiezen we θ (theta). We wijzen erop dat θ niet rechtstreeks observeerbaar is. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd[4]. De itemresponsfunctie $f_i(\theta)$ is gedefinieerd als een conditionele kans:

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

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

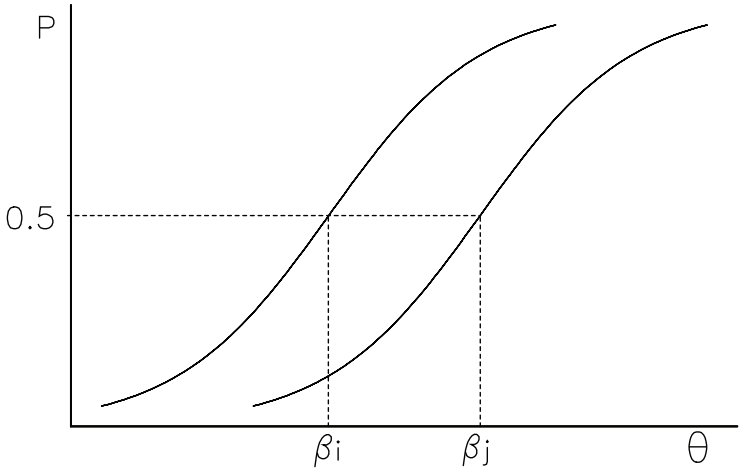
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \tag{2.2}$$

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 opgaven, *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, krijgen we:

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

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*. We kunnen de parameter β_i dus terecht omschrijven als de moeilijkheidsparameter van item *i*. De implicatie van het bovenstaande is dat 'moeilijkheid' en 'vaardigheid' op dezelfde schaal liggen.

Figuur 2.1 Twee itemresponscurven in het Raschmodel



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

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 opgave *j* juist te beantwoorden steeds kleiner is dan de kans op een juist antwoord op opgave *i*. Daaruit volgt de statistisch te toetsen voorspelling dat de verwachte proportie juiste antwoorden op opgave *j* kleiner is dan op opgave *i* in een willekeurige steekproef van personen. Splitst men nu een grote steekproef in twee deelsteekproeven, 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 blijkt echter lang niet altijd zo te zijn. Ook in ons geval niet. Veel van de opgaven blijken dan ook niet te kunnen worden beschreven met het Raschmodel. Daarom is bij dit instrument gekozen voor een ander IRT-model.

Alvorens het hier gebruikte model te introduceren, is een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattingsmethode 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 opgaven. 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 opgave *i*, gegeven de ruwe score, een functie is die alleen afhankelijk is van de itemparameters en onafhankelijk van de waarde van θ[5]. 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.

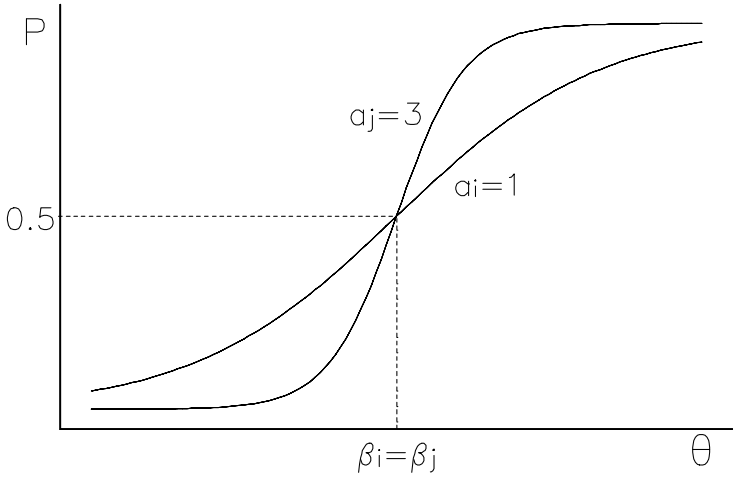
De CML-schattingsmethode is echter niet bij elk meetmodel toepasbaar. In het zogenaamde éénparameter logistisch model (One Parameter Logistic Model, afgekort: OPLM) is CML mogelijk. Dit model is, anders dan het Raschmodel, wel bestand tegen 'omwisseling' van 'proporties juist' in verschillende steekproeven (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 zogenaamde 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 opgaven *i* en *j*, die even moeilijk zijn maar verschillend discrimineren.

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

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



De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas en Verstralen, 1995). Dit programma voert eveneens 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 opgaven volgens het OPLM is dan ook een iteratief proces waarin alternerend de modelfit van de opgaven 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 gegevens-bestand. 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 opgaven in het normeringsonderzoek zijn meerkeuzeopgaven, 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 opgaven 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 meerkeuzeopgaven, 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 de populatieverdeling wordt gebruikgemaakt 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 die voor de schatting gebruikt wordt een aselecte steekproef uit de populatie is. In hoofdstuk 4 tonen we aan dat aan deze voorwaarde voldaan is. Daardoor is het mogelijk om voor elk normeringsmoment een schatting te maken van deze (normaal verdeelde) vaardigheidsverdeling.

3 Beschrijving van de toets

3.1 Opbouw, vorm, afname en rapportage

Het toetspakket Begrijpend luisteren voor groep 3 uit het Cito Volsysteem primair onderwijs bestaat uit twee toetsen per afnamemoment: een toets M3 die halverwege het schooljaar moet worden afgenomen (in januari/februari) en een toets E3 die aan het einde van het schooljaar (in mei/juni) moet worden afgenomen.

Opbouw
De toetsen voor groep 3 bestaan elk uit twee delen. Deze dienen bij voorkeur te worden afgenomen op twee verschillende dagdelen, zodat de leerlingen geconcentreerd aan de beide delen kunnen werken.
Elk deel bestaat uit 23 opgaven. De leerlingen maken 46 opgaven halverwege en 46 opgaven aan het einde van het schooljaar.

Vorm
De toetsen voor groep 3 bevatten een aantal korte luisterfragmenten met vragen. De leerlingen luisteren naar een luisterfragment met de bijbehorende vraag. Elk luisterfragment bestaat uit een of twee zinnen of uit een kort verhaaltje. Zowel de luisterfragmenten als de vragen worden voorgelezen en op cd aan de leerlingen aangeboden. Er is met andere woorden sprake van een ‘eenrichtingssituatie’.
De opgaven in de toetsen Begrijpend luisteren zijn meerkeuzeopgaven. Elke opgave bevat drie antwoord-alternatieven in de vorm van een illustratie, onder elke illustratie staat een antwoordhokje. We hebben ervoor gekozen om de leerlingen in groep 3 geen leesopgaven voor te leggen, omdat nog niet alle leerlingen in deze jaargroep het technisch lezen in voldoende mate beheersen. Geschreven antwoord-alternatieven zouden de prestaties van deze leerlingen negatief kunnen beïnvloeden en daarmee hun luistervaardigheid kunnen onderschatten.

Afname
De toetsen worden klassikaal afgenomen door de leerkracht, aan de hand van een afnamekaart met daarop afname-instructies. De afname start met een klassikale instructie en een aantal oefenopgaven.
De luisterfragmenten en de bijbehorende vragen staan op cd. Tijdens de afname bedient de leerkracht de cd-speler. De leerlingen luisteren naar luisterfragmenten die steeds worden gevolgd door een vraag. Direct na elke vraag volgt er een geluidssignaal (een piep) en zet de leerkracht de cd-speler op pauze.
De leerlingen hebben de tijd om over hun antwoord na te denken en hun antwoord in te vullen.
Zij beantwoorden de vragen door in hun opgavenboekje het antwoordhokje van hun keuze in te kleuren.

Rapportage
De toetsen Begrijpend luisteren zijn 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 bijlage van de handleiding 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 bespreken we in 3.2.1 de vaardigheidsaspecten die we aan de luistervaardigheid hebben onderscheiden. Daarnaast besteden we aandacht aan de verschillende opgaventypen die deze vaardigheidsaspecten representeren én geven we aan op welke vaardigheden deze aspecten een beroep doen. In 3.2.2 komen de criteria aan bod, zoals we die hebben gehanteerd bij het samenstellen van de toetsen Begrijpend luisteren.
De informatie in deze paragraaf vormt een aanvulling op de inhoudsverantwoording die is opgenomen in het toetspakket Begrijpend luisteren groep 3. (Zie hoofdstuk 6 Inhoudsverantwoording in de Handleiding.)

3.2.1 Begrijpend luisteren: een inhoudsanalyse

Voorafgaand aan de opgavenconstructie voor de toetsen Begrijpend luisteren hebben we ons de volgende vragen gesteld: Welke aspecten van de luistervaardigheid willen we in de toetsen onderbrengen? Welke opgaventypen representeren deze vaardigheidsaspecten? En op welke wijze verdelen we de opgaven over de diverse vaardigheidsaspecten en opgaventypen?

Vaardigheidsaspecten en opgaventypen
Op basis van de indeling van Krom, Ouborg en Kamphuis (2001) en in combinatie met hetgeen we in de literatuur hebben aangetroffen (cf. paragraaf 2.4.1), zijn we tot een aantal vaardigheidsaspecten gekomen die we onderscheiden aan de luistervaardigheid. Deze aspecten karakteriseren we als volgt:

- Het onthouden van gegeven informatie in gesproken taaluitingen: het geheugen zorgt voor het verwerken van alle informatie die bij de luisteraar binnenkomt. Horen wordt pas na inschakeling van het werkgeheugen luisteren. Het werkgeheugen bewerkt spraakklanken zodanig dat deze uiteindelijk resulteren in een samenhangende boodschap.
- Het doorzien van samenhang binnen of tussen gesproken taaluitingen of het afleiden van geïmpliceerde informatie: vaardige luisteraars herkennen in gesproken taal verbanden die de spreker al dan niet expliciet maakt. Ze worden daarbij geholpen door verbindingswoorden, grammatische stijlmiddelen en akoestische kenmerken die op deze verbanden wijzen.
- Het doorzien van de grammaticale organisatie van gesproken taaluitingen: om tot begrip van gesproken taal te komen, moeten luisteraars de grammaticale organisatie van allerlei typen uitingen vrijwel onmiddellijk kunnen doorzien. Belangrijke steunpilaren daarbij zijn de mogelijkheid tot anticiperen, vertrouwdheid met het taalsysteem en met de werking van akoestische kenmerken als intonatie, accentuatie en spreeksnelheid.
- Het doorzien van niet-letterlijk bedoeld taalgebruik in gesproken taaluitingen: om de juiste betekenis of boodschap aan gesproken taal toe te kunnen kennen, is het van belang dat luisteraars doorzien wanneer het de spreker niet gaat om de letterlijke betekenis van zijn taaluiting(en), maar om de achterliggende betekenis die hij via zijn woorden wil overbrengen.
- Het afleiden van geïmpliceerde informatie uit gesproken taaluitingen: kennis van de wereld is onmisbaar als het gaat om het verlenen van betekenis aan gesproken taal, waarin ‘niet alles’ gezegd wordt en de luisteraar ontbrekende informatie moet afleiden. Hij maakt in dat geval aanspraak op zijn kennis van de buitentalige werkelijkheid, zoals zijn kennis van het taalsysteem of zijn kennis over gewoonten en gebruiken.
- Het herkennen van de globale inhoud van de gesproken taaluitingen: luisteraars bouwen tijdens het luisteren in hun hoofd een representatie op van datgene wat ze horen. Ze brengen allerlei inhoudelijke elementen (personen, plaatsen of tijden) en de relaties en verbanden tussen die elementen (motieven, handelingen, resultaten) in hoofdlijnen ‘in kaart’. Dit resulteert in een interne representatie van de globale inhoud en structuur van een reeks van uitingen, waardoor de luisteraar in staat is de strekking van een ‘verhaal’ te begrijpen.

Bij elk vaardigheidsaspect zijn verschillende opgaventypen ontwikkeld. In onderstaand overzicht geven we per vaardigheidsaspect schematisch weer welke opgaventypen eronder vallen. Voor voorbeelden van concrete opgaven bij de diverse opgaventypen verwijzen we naar bijlage 1.

Het onthouden van gegeven informatie in gesproken taaluitingen
- opgaventype ‘combineren van inhoudselementen’
- opgaventype ‘selecteren van inhoudselementen’

Het doorzien van samenhang binnen of tussen gesproken taaluitingen of het afleiden van geïmpliceerde informatie
- opgaventype ‘expliciete verbanden’
- opgaventype ‘impliciete verbanden’
- opgaventype ‘anaforische verwijsrelaties’

Het doorzien van de grammaticale organisatie van gesproken taaluitingen
- opgaventype ‘grammaticale constructies’

Het doorzien van niet-letterlijk bedoeld taalgebruik in gesproken uitingen
- opgaventype ‘gebruik van metaforen’
- opgaventype ‘ironisch of sarcastisch taalgebruik’
- opgaventype ‘verrichten van taalhandelingen’

Het afleiden van geïmpliceerde informatie uit gesproken taaluitingen
- opgaventype ‘kennis van het taalsysteem’
- opgaventype ‘kennis over gewoonten en gebruiken’

Het herkennen van de globale inhoud van de gesproken taaluitingen
- opgaventype ‘globale betekenistoekenning’

Verdeling van de opgaven over de toetsen

Zoals vermeld in paragraaf 2.4.1 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. Echter, de vluchtige aard van gesproken taal maakt van reflectie in een eenrichtingssituatie een complexe aangelegenheid, waarmee nog maar weinig ervaring is opgedaan in de evaluatieve context waar het leerlingen in de onderbouw van het basisonderwijs betreft. We hebben er daarom voor gekozen om in de toetsen Begrijpend luisteren voor groep 3 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 aanspraak maken tijdens het luisteren naar gesproken taal.

Bij het samenstellen van de toetsen zijn we ervan uitgegaan dat de vaardigheden Begrijpen en Interpreteren in het luisterproces een vrijwel even belangrijke rol vervullen. We hebben daarom gestreefd naar een evenwichtige opgavenverdeling over deze vaardigheden, zowel per toets als over de beide toetsen heen.

Tabel 3.1 Aantal opgaven Begrijpen en Interpreteren in de toetsen Begrijpend luisteren groep 3

Toets	Aantal opgaven Begrijpen	Aantal opgaven Interpreteren	Totaal aantal opgaven
M3	23 (50%)	23 (50%)	46
E3	22 (48%)	24 (52%)	46

Tabel 3.1 laat zien dat (ongeveer) de helft van alle opgaven in de toetsen Begrijpend luisteren voor groep 3 een beroep doet op de vaardigheid Begrijpen en de andere helft op de vaardigheid Interpreteren. Dit komt nauw overeen met de fifty-fifty-verdeling die we beoogd hadden.

We hebben geprobeerd om de opgaven evenredig te verdelen over de diverse vaardigheidsaspecten. Dit verzekerde ons er tijdens de constructiefase van dat de luistervaardigheid in al haar facetten en van alle kanten belicht werd. Uiteindelijk zijn alle vaardigheidsaspecten in voldoende mate in de toetsen vertegenwoordigd. In werkelijkheid zijn de vaardigheidsaspecten echter niet zo duidelijk van elkaar te scheiden en grijpen ze op elkaar in, beïnvloeden ze elkaar en bouwen ze op elkaar voort. We kunnen ze dan ook niet opvatten als te isoleren aspecten en vaardigheden van het begrijpend luisteren. Het feit dat de opgaven op één vaardigheidsschaal liggen, illustreert dit ook. We zijn er daardoor zeker van dat het om één uni-dimensionele vaardigheid gaat (zie ook hoofdstuk 6).

Tabel 3.2 geeft de uiteindelijke verdeling weer van de opgaven over de onderscheiden vaardigheden en vaardigheidsaspecten.

Tabel 3.2 Verdeling van de opgaven naar vaardigheid en vaardigheidsaspecten per afnamemoment

Vaardigheid		**M3**	**E3**
Begrijpen			
Onthouden van gegeven informatie in gesproken taaluitingen		11 (46%)	11 (47%)
Doorzien van samenhang binnen of tussen gesproken taaluitingen (…)		12 (50%)	11 (48%)
Doorzien van de grammaticale organisatie van gesproken taaluitingen		1 (4%)	1 (4%)
	Totaal	*24* (100%)	*23* (100%)
Interpreteren			
Doorzien van niet-letterlijk bedoeld taalgebruik in gesproken taaluitingen		8 (36%)	3 (13%)
Afleiden van geïmpliceerde informatie uit gesproken taaluitingen		12 (55%)	17 (74%)
Herkennen van de globale inhoud van gesproken taaluitingen		2 (9%)	3 (13%)
	Totaal	*22 (100%)*	*23 (100%)*

3.2.2 Selectie van de opgaven

Alle opgaven die in de toetsen Begrijpend luisteren zijn opgenomen, zijn speciaal voor deze toetsen geconstrueerd door een constructieteam, voornamelijk bestaande uit (oud-)leerkrachten uit het basisonderwijs. Allereerst zijn er in een landelijk proefonderzoek opgaven voorgelegd aan leerlingen in groep 3, 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. Het gaat dan bijvoorbeeld om opgaven die vaker door vaardige leerlingen dan door minder vaardige leerlingen fout gemaakt worden. Daarnaast biedt een proefafname de mogelijkheid om aan de deelnemende leerkrachten te vragen of ze inhoudelijke of andersoortige bezwaren hebben tegen de luisterfragmenten of opgaven zoals die in de toetsen zijn opgenomen.
De opgaven die psychometrisch geschikt bleken, zijn vervolgens opgenomen in de toetsen 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 verschillende vaardigheids-aspecten en opgaventypen. Daarnaast probeerden we te vermijden dat opgaven die elkaar ‘bijten’ (bijvoorbeeld omdat de afbeeldingen op elkaar lijken) in hetzelfde deel van een van de toetsen geplaatst

zouden worden en zochten we naar een zekere balans in de lengte van de luisterfragmenten. Bij wijze van uitzondering hebben we wel eens een opgave gehandhaafd die wat te moeilijk of te gemakkelijk was of een opgave met een te laag discriminerend vermogen.
Van alle opgaven die zijn meegegaan in het normeringsonderzoek zijn de gekalibreerde p-waarde en de r_{ir} –waarde bepaald. De uiteindelijke selectie en verdeling van de opgaven over de toetsen was steeds een zo goed mogelijk compromis tussen de eisen op psychometrisch en inhoudelijk gebied en tussen overwegingen van praktische aard. Uiteindelijk zijn er 46 opgaven per afnamemoment in de toetsen opgenomen.

4 Het normeringsonderzoek

4.1 Opzet en verloop

Met het oog op het ontwikkelen van de toetsen Begrijpend luisteren zijn in 2009 voor de afnamemomenten medio groep 3 (M3) en eind groep 3 (E3) opgaven geconstrueerd.
In mei en juni 2010 zijn deze opgaven in een kalibratieonderzoek (proefonderzoek) voorgelegd aan groepen leerlingen 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 de normeringsmomenten medio en einde schooljaar. De normering voor de toets M3 vond plaats in januari en februari 2011, de normering voor de toets E3 in mei en juni 2011. De kalibratie- en normeringsonderzoeken samen maakten het mogelijk de ontwikkeling van de vaardigheid in begrijpend luisteren in kaart te brengen.

Het kalibratieonderzoek
Eerder merkten we al op dat in het kalibratieonderzoek is uitgegaan van een onvolledig design: niet alle leerlingen in de steekproef van het kalibratieonderzoek hebben alle opgaven gemaakt. De opgaven werden verdeeld over clusters en aan elke leerling werden meerdere opgavenclusters voorgelegd. Clusters die gezamenlijk aan een groep leerlingen worden voorgelegd, worden 'boekjes' genoemd; de verschillende boekjes overlappen elkaar. Deze overlap zorgt ervoor dat het design verbonden is, een noodzakelijke voorwaarde om CML-schattingen van de itemparameters (CML-estimates) te kunnen bepalen. Een voorbeeld van zo'n design staat in de verantwoording van de toetsen Begrijpend lezen (Staphorsius, Krom, Kleintjes en Verhelst, 2004).
In het kalibratieonderzoek zijn de opgaven begrijpend luisteren voorgelegd aan 1450 leerlingen in groep 3. De opgaven waren verdeeld over zes verschillende boekjes in een onvolledig maar 'verbonden' design. In tabel 4.1 is te zien hoe dit design eruitzag. Elk boekje bevatte gemiddeld 58 opgaven, bedoeld voor de afnamemomenten M3 en E3. Elke opgave is door gemiddeld 484 leerlingen beantwoord.

Tabel 4.1 Afnamedesign kalibratieonderzoek (proefonderzoek) groep 3

Boekje	Taak1	Taak2	Taak3	Taak4	Taak5	Taak6
3.1	ME3	ME3				
3.2		ME3	ME3			
3.3			ME3	ME3		
3.4				ME3	ME3	
3.5					ME3	ME3
3.6	ME3					ME3

Het normeringsonderzoek
Het normeringsonderzoek levert aanvullende gegevens op over de kwaliteit en de moeilijkheid van de opgaven en over de landelijke verdeling van de vaardigheid van de leerlingen op de verschillende afnamemomenten. Tijdens dit onderzoek zijn de leerlingen getoetst om in een landelijke normgroep referentiegegevens voor de verschillende afnamemomenten te kunnen verzamelen om op basis daarvan de ontwikkeling van de vaardigheid in begrijpend luisteren in kaart te kunnen brengen. In groep 3 is een deel van de leerlingen gevolgd; deze leerlingen hebben zowel meegedaan aan het normeringsonderzoek M3 als aan het normeringsonderzoek E3.
De gegevens uit de normeringsonderzoeken zijn ingezet om de vaardigheidsverdelingen op de verschillende normeringsmomenten (M3-januari/februari en E3-mei/juni) te kunnen bepalen.

De resultaten van de leerlingen die hebben deelgenomen aan de normeringsonderzoeken, vormen de basis voor de normeringsgegevens zoals die zijn opgenomen in de handleiding bij de toetsen Begrijpend luisteren: de gemiddelden en de standaardafwijking voor de afzonderlijke afnamemomenten (zie tabel 4.9). De representativiteit van deze leerlingen wordt in paragraaf 4.2 voor elk van de normeringsonderzoeken weergegeven op basis van schoolkenmerken. Het opnemen van een anker met het vorige normeringsmoment was hierbij wenselijk. Voor alle normeringsmomenten is het totaal aantal leerlingen ruim voldoende om een verantwoorde normering op te baseren, zoals te zien is in tabel 4.2.
In tabel 4.2 worden de aantallen leerlingen per afnamemoment gegeven. Leerlingen die meer dan een kwart van de opgaven niet gemaakt hebben of waarbij een taak ontbrak, zijn niet in de berekening van de normeringsgegevens meegenomen, omdat het onduidelijk is of ze serieus aan de toets gewerkt hebben. Bij zowel M3 als E3 kwamen geen hoge percentages ontbrekende antwoorden voor, dus het verschil tussen ‘deelname’ en ‘onderzoek’ wordt hier alleen veroorzaakt door ontbrekende taken. Zouden deze leerlingen in de berekening betrokken worden dan zouden ze de normeringsresultaten kunnen ‘vervuilen’. De uiteindelijke aantallen leerlingen die in de normeringsanalyses zijn opgenomen, staan in de kolom ‘Onderzoek’.

Tabel 4.2 Aantal leerlingen per afnamemoment

Afnamemoment	Aantal leerlingen	
	Deelname	Onderzoek
M3	1198	1192
E3	2063	2037

4.2 Representativiteit

In deze paragraaf bespreken we de representativiteit van de normeringssteekproef voor groep 3. Voor de afnamemomenten M3 en E3 zijn normeringsonderzoeken uitgevoerd. Voor deze onderzoeken werden steekproeven getrokken van scholen; deze scholen werd gevraagd deel te nemen met hun groep(en) 3 aan het onderzoek. Bij de steekproeftrekking werd rekening gehouden met de representativiteit van de scholen op basis van de volgende schoolkenmerken: percentage achterstandsleerlingen, geografische spreiding en mate van verstedelijking.
Naar aanleiding van de onderzoeken beschikken we over de achtergrondkenmerken van deze scholen. We beschrijven hieronder de representativiteit van de scholen waar de leerlingen uit tabel 4.2 uit de vorige paragraaf onderwijs volgen.

Percentage achterstandsleerlingen (stratum)
Een belangrijk schoolkenmerk is het percentage achterstandsleerlingen. Om hier zicht op te krijgen heeft Cito het begrip ‘stratum’ ingevoerd, een concept waarbij gebruik is gemaakt van de CFI-leerlinggewichten. De populatiegegevens zijn ontleend aan CFI-gegevens voor 6956 scholen in 2011. Cito hanteert hierbij vier strata (zie tabel 4.3) waarbij het gaat om het percentage achterstandsleerlingen (gebaseerd op leerlinggewichten, zie tabel 4.3a). De omschrijving geeft een inhoudelijke interpretatie van de strata: stratum 1 bevat de scholen met meer dan 40% achterstandsleerlingen, stratum 2 de scholen met 25-40% achterstandsleerlingen, stratum 3 de scholen met 10-25% achterstandsleerlingen en stratum 4 de scholen die minder dan 10% achterstandsleerlingen bevatten. Uit PPON-onderzoeken blijkt stratum een belangrijke verklarende variabele voor schoolprestaties. Belangrijk om op te merken is dat de leerlinggewichten – ook wel formatiegewichten genoemd – gebaseerd zijn op de opleiding van de ouders (zie tabel 4.3b). Herkomst is in de huidige regeling (Besluit 283, Wijziging Besluit bekostiging WPO in verband met wijziging gewichtenregelingen, 2006) niet opgenomen. Een omschrijving van de gebruikte opleidingsniveaus is te vinden in tabel 4.3b.

Tabel 4.3 Scholen uit de steekproef (normeringsonderzoek verdeeld naar stratum

	Populatie			Steekproef		
Stratum	**Aantal**	**%**	**M3**	**%**	**M3**	**%**
1	493	7,0	5	10,5	11	12,5
2	530	7,6	7	14,6	12	13,7
3	1987	28,5	17	35,4	27	30,7
4	3955	56,8	19	39,6	38	43,1
Totaal	***6965***	*100,0*	*48*	*100,0*	*88*	*100,0*

Tabel 4.3a Gewichtenregeling

Gewicht	**Omschrijving**
0.3	Beide ouders of de ouder die belast is met de dagelijkse verzorging heeft een opleiding uit categorie 2 gehad
1.2	Eén van de ouders heeft een opleiding gehad uit categorie 1 en de ander een opleiding uit categorie 1 óf 2
0	Eén van de ouders of beide ouders hebben een opleiding gehad uit categorie 3

Tabel 4.3b Opleidingsniveau ouder

Categorie	**Omschrijving**
1	Maximaal basisonderwijs of (v)so-zmlk
2	Maximaal lbo/vbo, praktijkonderwijs of vmbo basis- of kaderberoepsgerichte leerweg
3	Overig vo en hoger

In tabel 4.3 is de verdeling van de scholen uit de steekproef voor het normeringsonderzoek naar stratum weergegeven. Een vergelijking van de steekproefverdeling met de landelijke verdeling laat zien dat voor de twee afnamemomenten de normeringssteekproef representatief genoemd mag worden. Stratum 1, 2 en 3 zijn licht oververtegenwoordigd en stratum 4 is enigszins ondervertegenwoordigd in de steekproeven. Om die reden is onderzocht of er wellicht een weging zou moeten plaatsvinden. Het onderzoek naar die weging is samengevat in tabel 4.4 en 4.4a. Met het programma Saul zijn de gemiddelden en standaardafwijkingen van de steekproeven M3 en E3 voor de vier strata berekend. Vervolgens zijn uit de resulterende verdelingen afnamen gesimuleerd, proportioneel naar de populatieverdeling van de scholen over de strata. Van de resulterende teruggewogen populaties worden de gemiddelden en standaarddeviaties weergegeven in tabel 4.4a. Als deze gemiddelden vergeleken worden met de gemiddelden van de steekproeven in tabel 4.4 (onderaan bij 'totaal'), dan is te zien dat dit verschil zeer klein is. Daarom is besloten geen weging toe te passen.

Tabel 4.4 Gemiddelden en standaardafwijking per stratum (gegevens zonder weging)

	Marginalen M3		Marginalen E3	
Stratum	**M**	**SD**	**M**	**SD**
1	37,8	5,98	41,6	8,09
2	45,3	8,92	47,2	8,72
3	47,9	8,19	51,5	8,39
4	48,2	8,78	51,9	8,85
totaal	47,6	8,72	50,1	8,68

Tabel 4.4a Gemiddelden en standaardafwijkingen van gesimuleerde afnamen teruggewogen naar stratumverdeling in de populatie

Afname-moment	Aantal gesimuleerd	Gemiddelde vaardigheid	Standaard-afwijking
M3	69650	47,7	8,72
E3	69650	50,3	8,68

Representativiteit naar geografische spreiding

De verdeling van alle scholen in Nederland en van de scholen in de normeringssteekproef naar regio is te vinden in tabel 4.5. Regio Noord bevat de provincies Groningen, Friesland en Drenthe; regio Oost de provincies Overijssel, Gelderland en Flevoland; regio West de provincies Utrecht, Noord-Holland en Zuid-Holland en Zeeland en regio Zuid bestaat uit Noord-Brabant en Limburg. De steekproefverdeling wijkt voor de afnamemomenten nauwelijks af van de populatieverdeling. Dit biedt steun voor de bewering dat de scholen in de normeringssteekproef representatief naar regio zijn.

Tabel 4.5 Verdeling scholen naar regio per afnamemoment in aantallen

Regio	Landelijk		Steekproef			
	Aantal	%	M3	%	E3	%
Noord	1092	16	6	10%	15	17%
Oost	1705	24	15	29%	22	25%
West	2880	41	20	47%	33	38%
Zuid	1288	19	7	14%	18	20%
Totaal	*6965*	*100*	*48*	*100%*	*88*	*100%*

Representativiteit naar verstedelijking

De verdeling naar verstedelijking van alle scholen en van de scholen in de normeringssteekproeven wordt weergegeven in tabel 4.6. Voor verstedelijking komen de verdelingen in de steekproeven bijna exact overeen met de verdeling in de populatie (Chi-kwadraat M3 = 0.026, df = 1, p = 0.87; Chi-kwadraat E3 = 0.48, df = 1, p = 0.48). Dit biedt steun voor de bewering dat de scholen in de normeringssteekproef representatief naar verstedelijking zijn.

Tabel 4.6 Aantal scholen naar verstedelijking per afnamemoment in aantallen

Mate van verstedelijking	Landelijk		Steekproef			
	Aantal	%	M3	%	E3	%
Stedelijk	2370	34	44	33%	33	38%
Landelijk	4593	66	88	67%	55	62%
Totaal	*6963*	*100*	*132*	*100%*	*88*	*100%*

Representativiteit naar sekse en leeftijd

Representativiteit naar sekse en representativiteit naar leeftijd zijn niet onderzocht. Omdat steeds volledige klassen hebben deelgenomen aan het normeringsonderzoek, is er geen aanleiding om te veronderstellen dat de verdeling van meisjes en jongens en de verdeling naar leeftijd in de normgroep afwijkt van de verdeling zoals gebruikelijk in leerjaar 3.

4.3 Kalibratie en normering

4.3.1 Resultaten kalibratie- en normeringsonderzoek

In de inleiding merkten we al op dat in het kalibratieonderzoek dat aan de opgavenbanken ten grondslag ligt, is uitgegaan van een onvolledig design: niet alle leerlingen in de steekproef van het kalibratieonderzoek maken alle opgaven.
Na de kalibratie in de normeringsfase vormen de opgaven een gekalibreerde opgavenbank. Bij de analyse van de leerlingantwoorden is nagegaan of de verschillende opgaven en opgaventypen een beroep doen op hetzelfde complex aan vaardigheden. Dit bleek het geval te zijn. Opgaven die niet voldeden aan de passingscriteria die we hierna beschrijven, zijn uit de verzameling verwijderd.

4.3.2 Stappen in de kalibratie en toetsing van het IRT-model

Met kalibratie wordt bedoeld dat we kengetallen zoeken bij de items 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) en 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 kalibratieprocedure. De statistische toetsen in OPLM hebben goede statistische en asymptotische eigenschappen daar OPLM behoort tot de exponentiële familie, met de gewogen somscore:

$$s = \sum_{i=1}^{k} a_i x_i \quad , \qquad (4.1)$$

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:

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

Deze zogenaamde 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:

$$S = f(p(+\mid s) - prop(+\mid s)) \quad . \qquad (4.3)$$

Deze zogenaamde S-toets heeft een chi-kwadraatverdeling onder het model. Analoog hieraan is er ook een toets om vormen van vraagonzuiverheid (in het Engels 'item bias' of 'differential item functioning', afgekort DIF) op te sporen:

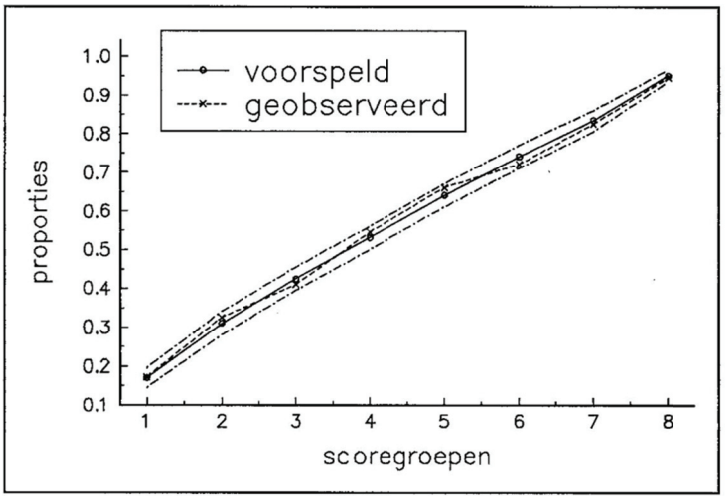
$$S = h(p_I(+\mid s) - prop_I(+\mid s), (p_{II}(+\mid s) - prop_{II}(+\mid s)) \quad (4.4)$$

waarbij I en II de twee niveaus van de variabele indiceren waarvoor we de bias onderzoeken.
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 Kies geschikte waarden voor de discriminatie-indices in OPLM.
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. Zie hiervoor tabel 4.7a, 4.7b, 4.8 en figuur 4.2a en 4.2b en tabel 4.8. Afhankelijk van de uitkomsten kunnen items worden verwijderd. Ook inhoudelijke overwegingen (zie hiervoor hoofdstuk 2 en 3 over de achtergronden van de toetsinhoud) spelen een rol in dit beslissingsproces.
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 4.1 (zie Staphorsius, 1994, blz. 239). Figuur 4.1 beeldt voor een opgave de gegevens af waarop de betreffende S -toetsen gebaseerd zijn (zie handleiding OPLM: Verhelst, 1992). 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 grote 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 parameter-schattingen 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).

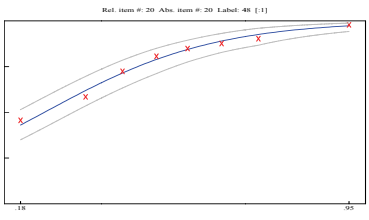
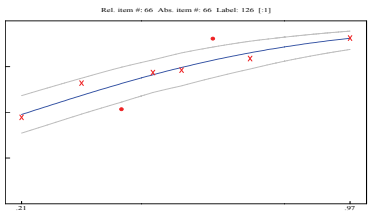
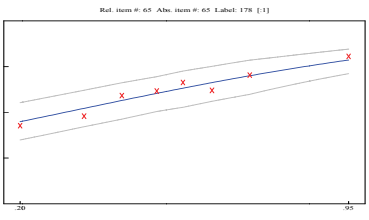
Figuur 4.1 Grafische voorstelling van een S –toets



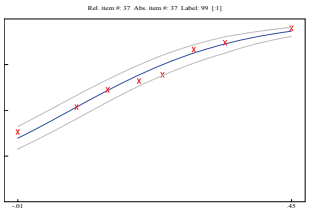
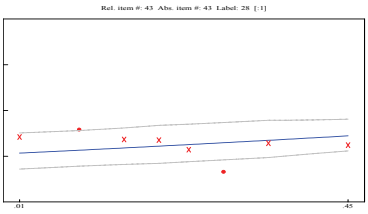
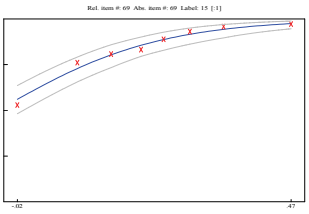
Het is ondoenlijk om voor alle in de toetsen opgenomen items een grafische weergave van de S-toets op te nemen, zoals die in figuur 4.1. We volstaan daarom voor beide toetsen met enkele voorbeelden (zie figuur 4.2a voor M3 en 4.2b voor E3). Daarbij zijn vooral de voorbeelden die een representatieve passing en de slechtste passing laten zien interessant. De afbeeldingen voor opgaven met een representatieve passing staan model voor de overgrote meerderheid van de toetsopgaven. Dat impliceert dat bij de meeste items alle geobserveerde proporties binnen het 95%-betrouwbaarheidsinterval van de geschatte proporties liggen. Bij enkele items is dit niet het geval, maar aan de afbeeldingen voor het slechtst passend item is te zien dat slechts twee van de acht geobserveerde proporties buiten dit interval vallen. Dit impliceert dat ook voor deze items de model-fit nog zeer behoorlijk is.
Dit is ook te zien aan de verdeling van de rechter overschrijdingskansen van alle uitgevoerde S-toetsen (zie tabel 4.7a en 4.7b). Deze zijn vrij uniform verdeeld over het gehele [0,1]-interval met een beperkt aantal significante waarden. Op basis van deze gegevens kunnen we concluderen dat de model-fit op het niveau van de individuele items uitstekend is.

Ook op het niveau van de toets als geheel is er sprake van een goede fit, wat blijkt uit de R1c-waarden voor beide toetsversies (zie tabel 4.8). Deze voldoen aan de vuistregel dat R1c niet hoger zou moeten zijn dan anderhalf maal het aantal vrijheidsgraden.

Figuur 4.2a Voorbeelden van S-toetsen voor Begrijpend luisteren M3 met de best passende, de slechtst passende en een qua passing representatieve opgave

		
Best passend	Slechtst passend	Representatieve passing

Figuur 4.2b Voorbeelden van S-toetsen voor de E3 Begrijpend luisteren met de best passende, de slechtst passende en een qua passing representatieve opgave

		
Best passend	Slechtst passend	Representatieve passing

Tabel 4.7a Verdeling van overschrijdingskansen bij S-toetsen voor M3

0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
0	5	8	7	7	8	6	9	8	13	8	11

Tabel 4.7b Verdeling van overschrijdingskansen bij S-toetsen voor E3

0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
2	2	3	7	10	13	5	11	11	13	16	12

Tabel 4.8 R1c-waarden

	R1c	**df**	**p**
M3	722.92	619	0.00
E3	835.211	604	0.00

Voor de toetsen M3 en E3 Begrijpend luisteren is ook de zogeheten constante ‘c’ berekend waarnaar in het beoordelingssysteem van de Cotan wordt verwezen (Evers et al., 2010). Daaruit komt naar voren dat in de toets M3 77% van de opgaven een c ≤ 0,2 en daarmee het predicaat ‘goed’ krijgt (volgens de criteria van het Cotan-beoordelingssysteem). 23% van de opgaven in de toets M3 heeft een constante ‘c’ die tussen 0,3 en 0,4 ligt en krijgt volgens diezelfde criteria het predicaat ‘voldoende’. Bij E3 kon 99% van de opgaven op deze manier als ‘goed’ gekwalificeerd worden en 1% als ‘voldoende’. In beide toetsen komen dus geen opgaven voor met een als onvoldoende te kenschetsen c-waarde.

Al deze resultaten wijzen erop dat we kunnen spreken van een geslaagde kalibratie. Hiermee is het laatste woord nog niet gezegd over de validiteit, maar het kalibratieonderzoek 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 kalibratieanalyse als puur formeel proces (het analyseren van een grote onvolledige tabel met nullen en enen) 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 opgaven in de opgavenbank meten dekkend is voor en samenvalt met het construct ‘begrijpend luisteren’ zoals in het didactisch en wetenschappelijk forum wordt bedoeld? De vraag is dan in het geval van begrijpend luisteren: kan het unidimensionele concept onder de opgaven in de opgavenbank Begrijpend luisteren inderdaad worden opgevat als een vaardigheid in

begrijpend luisteren? Voor een antwoord op deze vraag verwijzen we naar het hoofdstuk Validiteit (Hoofdstuk 6).

4.3.3 Normering

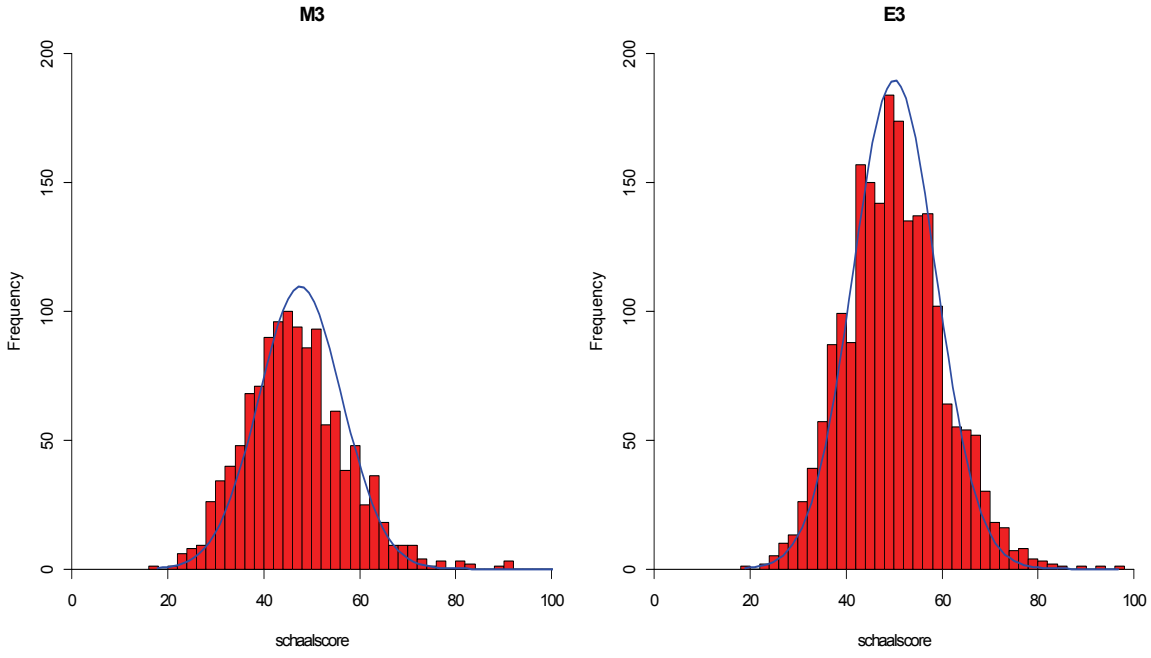
In paragraaf 2.4.2 gaven we belangrijke implicaties voor een gekalibreerde opgavenverzameling. Het slagen van de kalibratie betekent dat we met een selectie van opgaven uit de opgavenbank de vaardigheid bij een leerling kunnen meten. Hoe nauwkeurig we dat doen staat beschreven in paragraaf 5.2.
We kunnen nu een schatting maken van de verdelingen van de vaardigheid in een welomschreven populatie, omdat we opgaven (in boekjes) voorgelegd hebben aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. We schatten het gemiddelde en de standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met deze schattingen 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.
Een overzicht van de geschatte gemiddelden en de standaardafwijkingen van de vaardigheid op de verschillende normeringsmomenten is te vinden in tabel 4.9. Uit deze tabel blijkt dat de gemiddelde vaardigheid in begrijpend luisteren toeneemt, terwijl de spreiding nagenoeg gelijk blijft.

Tabel 4.9 Overzicht van de vaardigheidsverdelingen per afnamemoment

Afnamemoment	Aantal leerlingen	Gemiddelde vaardigheid	Standaardafwijking
M3	1198	47,63	8,72
E3	2037	50,08	8,68

Om een indruk te krijgen van de verdeling van de vaardigheid op de verschillende normeringsmomenten is de verdeling van de geobserveerde en de geschatte vaardigheidsverdeling voor elk van de normerings-momenten grafisch weergegeven in figuur 4.3

*Figuur 4.3 Geobserveerde en geschatte vaardigheidsverdeling voor de normeringsmomenten M3 en E3**



* De histogrammen zijn gebaseerd op verschillende aantallen leerlingen en daardoor moeilijk vergelijkbaar.

Op de horizontale as is de vaardigheid weergegeven; op de verticale as zien we de verdeling van de geobserveerde vaardigheid per afnamemoment. De ingetekende curve is steeds de geschatte vaardigheidsverdeling voor het betreffende afnamemoment. Het moge duidelijk zijn dat de vaardigheid goed beschreven kan worden met een normale verdeling[6].
We kunnen ook op een andere manier een antwoord krijgen op de vraag of de aanname van normaliteit wel terecht is geweest. Daartoe is in tabel 4.10 per afnamemoment de geobserveerde verdeling in de verschillende niveaus weergegeven. 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%). De indeling I tot en met V is symmetrisch opgebouwd: vijf niveaugroepen van ieder 20% (zie ook paragraaf 2.3).

Tabel 4.10 Geobserveerde verdeling van de niveaus per afnamemoment in percentages

	M3	**E3**
E	10,2	10,1
D	15,4	14,9
C	25,4	25,4
B	24,0	24,8
A	25,0	24,8
V	20,1	20,6
IV	20,1	19,8
III	20,4	20,4
II	21,5	19,4
I	17,9	19,7

Voor alle afnamemomenten en voor alle niveaus is te zien dat de afwijking tussen de geobserveerde en de theoretische percentages klein is. Uit bovenstaande blijkt dat de aanname van een normaal verdeelde vaardigheidsverdeling door de data ondersteund wordt.

[6] De histogrammen zijn gebaseerd op verschillende aantallen leerlingen en daardoor moeilijk vergelijkbaar.

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Betrouwbaarheid

In hoofdstuk 4 is onder meer aangegeven dat elke leerling die deelgenomen heeft aan het normerings-onderzoek slechts een deel van de opgaven gemaakt heeft die uiteindelijk in de toetsen Begrijpend luisteren opgenomen zijn. De betrouwbaarheid van de toetsen in klassieke zin is dan ook niet rechtstreeks te bepalen. Het is echter wel mogelijk om de betrouwbaarheid van iedere toets te schatten door gebruik te maken van het feit dat alle opgaven die zijn opgenomen in de toetsen OPLM-geschaald zijn. Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het feit dat de toetsen volledig bestaan uit OPLM-gekalibreerde opgaven. Om relevante beschrijvende gegevens bij de verschillende toetsen 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 de betrouwbaarheidscoëfficiënt uit de klassieke toetstheorie. 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 duiden we aan met τ(θ). Als we bovendien weten hoe θ in de populatie verdeeld is, kunnen we ook het gemiddelde en de variantie van de ware scores in de populatie bepalen. De variantie van de ware scores in de populatie duiden we aan 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 we met een eenmalige toetsafname niet meer zonder fout de waarde van θ kunnen bepalen. De variantie van de geobserveerde toetsscore duiden we aan met Var(t|τ(θ)), en door weer gebruik te maken van de distributie van θ in de populatie kunnen we ook de gemiddelde variantie van de geobserveerde toetsscores gaan berekenen.

$$Var(t) = E[Var(t \mid \tau(\theta))] \qquad (5.1)$$

Deze variantie kunnen we opvatten als de (gemiddelde) meetfoutvariantie in de metriek van de geobserveerde scores t. In analogie met de theorie over de betrouwbaarheid definiëren we dan

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

waarin MAcc staat voor 'Accuracy of Measurement'.

Tabel 5.1 bevat informatie over de meeteigenschappen van de vaardigheidsschaal Begrijpend luisteren. In de eerste kolom staat het afnamemoment. De maximumscore voor iedere toets is gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. De derde en vierde kolom geven de geschatte gemiddelde scores en standaarddeviaties van de leerlingen op de verschillende toetsen. De vijfde kolom bevat informatie over de geschatte standaardmeetfout van iedere toets. De laatste kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de verschillende toetsen is.

De betrouwbaarheidscoëfficiënten zijn zonder uitzondering voldoende hoog. Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen Begrijpend luisteren) geeft de COTAN aan dat een betrouwbaarheidscoëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 voldoende en een betrouwbaarheidscoëfficiënt hoger dan 0,80 goed (Evers et al., 2010). Op grond van dit criterium is de meetnauwkeurigheid van alle toetsen voldoende te noemen.

Tabel 5.1 Beschrijvende gegevens bij de toetsen Begrijpend luisteren

Toets	Aantal items	Gemiddelde	Standaard-deviatie	Standaard-meetfout	Betrouwbaarheid
M3	46	47,63	8,72	4,27	0,76
E3	46	50,08	8,68	4,59	0,72

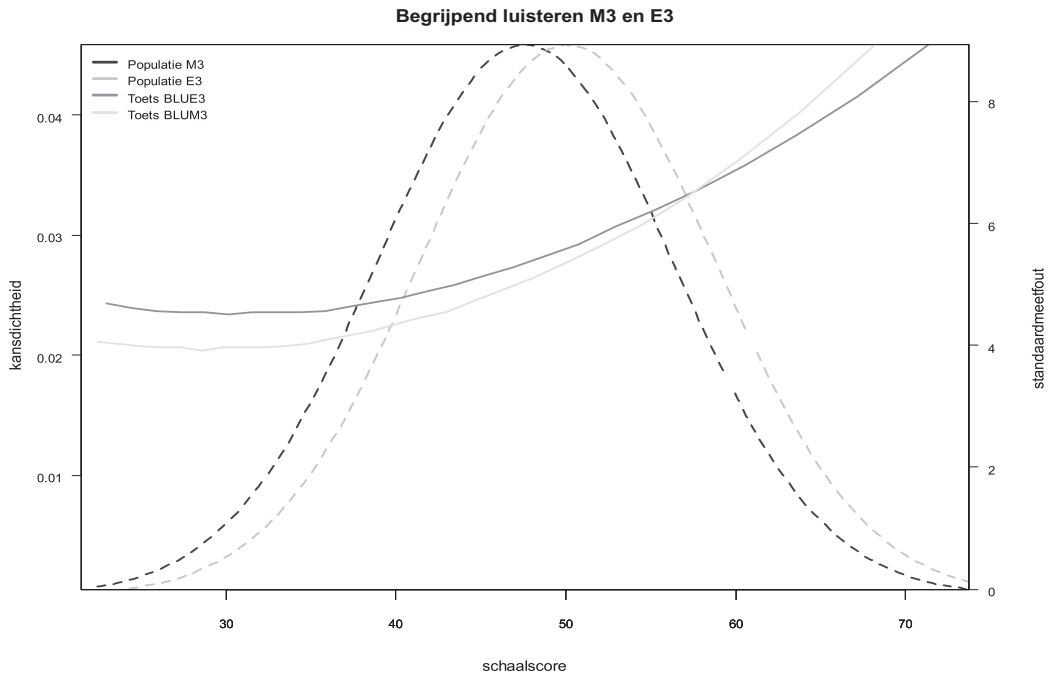
5.2 Meetnauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen Begrijpend luisteren en geven geen beeld van de lokale meetnauwkeurigheid. Figuur 5.1 doet dat wel.

Figuur 5.1 geeft grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de toetsen. In deze figuren is voor iedere toets de grootte van de meetfout afgebeeld. Ook zijn de kansdichtheidfuncties voor de normgroepen op de verschillende afnamemomenten opgenomen. Deze laten zien hoe de vaardigheid van de leerlingen verdeeld is over de vaardigheidsschaal in de populatie die de toets gemaakt heeft. De figuur maakt duidelijk dat de meetfout kleiner is in de lagere en de gemiddelde vaardigheidsniveaus dan in de hogere vaardigheidsniveaus. Het is de bedoeling dat het discriminerend vermogen van de toets vooral bij de zwakke leerlingen optimaal zou moeten zijn, omdat we met name de vaardigheid van deze leerlingen goed in kaart willen brengen.

Verdere gedetailleerde informatie over de meetnauwkeurigheid van de toetsen is te vinden in de handleiding van het toetspakket Begrijpend luisteren (Cito, 2011). Zie hiervoor de schaalscoretabellen in bijlage 2 van de handleiding waarbij in de laatste kolom het score-interval vermeld is. In deze kolom staat voor iedere ruwe score op elke toets het 67-procents-betrouwbaarheidsinterval voor de bijbehorende vaardigheidsschatting.

Figuur 5.1 Grootte van de meetfouten voor de toetsen M3 en E3 en de kansdichtheidfuncties voor de M3- en E3-populatie



6 Validiteit

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 aan de basis van de ontwikkeling van de opgaven de indeling in vaardigheidsaspecten ligt. Uiteindelijk zijn de diverse vaardigheidsaspecten in voldoende mate in de toetsen vertegenwoordigd. In werkelijkheid zijn de vaardigheidsaspecten echter niet zo duidelijk van elkaar te scheiden en grijpen ze op elkaar in, beïnvloeden ze elkaar en bouwen ze op elkaar voort. We kunnen ze dan ook niet opvatten als te isoleren aspecten en vaardigheden van het begrijpend luisteren. Het feit dat de opgaven op één vaardigheidsschaal liggen, illustreert dit ook. We zijn er daardoor zeker van dat het om één uni-dimensionele vaardigheid gaat. Een verdere aanwijzing voor de inhoudsvaliditeit is het gegeven dat (vrijwel) de helft van de opgaven een beroep doet op de vaardigheid Begrijpen en de andere helft 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, 2009, Krom e.a., 2011 en Sijtstra, 2005). In de toetsen Begrijpend luisteren zijn overigens 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. Bij het samenstellen van de toetsen zijn we uitgegaan van een evenwichtige opgavenverdeling over deze vaardigheden om de luistervaardigheid in al haar facetten en van alle kanten te kunnen belichten. Dit heeft uiteindelijk geresulteerd in een fifty-fifty-verdeling.

6.2 Begripsvaliditeit

De begripsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de toetsscores toe te schrijven zijn aan verklarende concepten en constructen die deel uitmaken van het theoretische kader dat aan de ontwikkeling van de toets ten grondslag ligt. Hieronder worden vier aanwijzingen voor de begripsvaliditeit van de toetsen Begrijpend luisteren beschreven. Deze hebben betrekking op de passing van het meetmodel (6.2.1), de convergente en de discriminante validiteit (6.2.2), de samenhang met de variabele leerjaar (6.2.3), de responsiviteit en de stabiliteit (6.2.4) en de itemkarakteristieken (6.2.5)

6.2.1 Passing van het meetmodel

De opgaven vormen na de kalibratie een gekalibreerde opgavenbank. Bij de analyse van de leerlingantwoorden is nagegaan of de verschillende opgaven een beroep doen op hetzelfde complex aan vaardigheden. Opgaven die niet voldeden aan de passingscriteria die we beschreven in paragraaf 4.3.2, zijn uit de opgavenverzameling verwijderd. Het betreft opgaven waarop werd gegokt, opgaven die onjuist geformuleerd zijn, opgaven die een slecht onderscheidend vermogen bleken te hebben, of opgaven die bij nader inzien toch niet alleen de vaardigheid in begrijpend luisteren bleken te meten.
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 geslaagde kalibratie maakt duidelijk dat het aannemelijk is dat er sprake is van unidimensionaliteit en dat deze gekalibreerde opgavenbank de latente trek meet die we de vaardigheid begrijpend luisteren noemen.

6.2.2 Convergente en discriminante validiteit

De convergente en discriminante validiteit van begrijpend luisteren kunnen worden onderzocht door middel van correlationeel onderzoek. Scholen die hebben deelgenomen aan het normeringsonderzoek E3 voor Begrijpend luisteren (Blui) is gevraagd om gegevens beschikbaar te stellen voor de vaardigheden Rekenen-Wiskunde (Rekwis; Cito, 2006), Spelling (Spel; Cito, 2006), Woordenschat (Wsch; Cito, 2009), Technisch lezen-Leestechniek (Tlez; Cito, 2009) en Technisch lezen-Leestempo (Ltemp; Cito, 2009). Daarnaast is gevraagd de startmodule van de toets Begrijpend lezen (Blez, Cito, 2006) af te nemen. De afnames hebben plaatsgevonden in mei/juni 2011, op het E-moment van groep 3.
Onze verwachting was dat de samenhang tussen begrijpend luisteren en technisch lezen matig zou zijn en dat de samenhang tussen de semantische onderdelen begrijpend lezen en begrijpend luisteren en tussen woordenschat en begrijpend luisteren groot zou zijn.

In tabel 6.1 worden de correlatiecoëfficiënten gerapporteerd tussen de hierboven genoemde toetsen. De correlatiecoëfficiënten zijn gecorrigeerd voor attenuatie. Als een toets bestaat uit een startmodule en een tweetal vervolgmodules (de toetsen Spelling en Technisch lezen) is ervoor gekozen om de laagste betrouwbaarheid aan te houden bij het bepalen van de attenuatie. In die gevallen is de gerapporteerde waarde de minimale waarde voor de correlatiecoëfficiënt.

Tabel 6.1 Correlaties tussen Begrijpend luisteren en andere toetsen

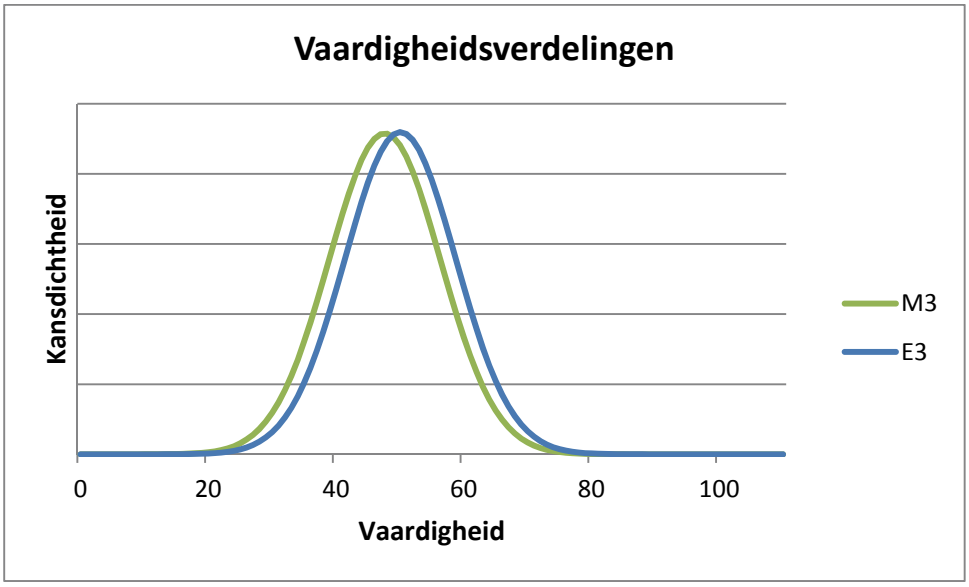
	Blui	**Wsch**	**Blez**	**Tlez**	**Ltemp**	**Spel**	**Rekwis**
Blui	1,00	0,77	0,51	0,41	0,40	0,38	0,44
Wsch		1,00	0,51	0,45	0,42	0,34	0,35
Blez			1,00	0,69	0,66	0,57	0,45
Tlez				1,00	0,80	0,67	0,45
Ltemp					1,00	0,56	0,35
Spel						1,00	0,50
Rekwis							1,00

Uit tabel 6.1 blijkt dat de correlaties tussen de toetsen Begrijpend luisteren (Blui) en de andere toetsen zoals Begrijpend lezen en Woordenschat hoger zijn dan tussen de toetsen die de meer technische taalonderdelen als Spelling, Technisch lezen en Leestempo-toetsen. Ook tussen Begrijpend luisteren en Rekenen-Wiskunde is de correlatie lager dan tussen Begrijpend luisteren en de onderdelen waarin semantische vaardigheden domineren. Dat de correlatie tussen de toetsen Begrijpend luisteren en Rekenen-Wiskunde iets hoger is dan de correlatie tussen Begrijpend luisteren en Spelling en tussen Begrijpend luisteren en de beide toetsen voor Technisch lezen kan verklaard worden uit het feit dat bij Rekenen-Wiskunde de opgaven redelijk talig zijn en in die zin ook bij Rekenen-Wiskunde het talig aspect een rol speelt.
Een aanwijzing voor de validiteit van de opgaven Begrijpend luisteren is de relatief hoge correlatie tussen Begrijpend luisteren en Woordenschat (r = 0,77). Woordenschat is dan ook een belangrijke ondersteunende vaardigheid bij Begrijpend luisteren (zie ook hoofdstuk 2).
Zoals verwacht is de correlatie tussen de toetsen Begrijpend luisteren en Begrijpend lezen hoger dan de correlatie tussen Begrijpend luisteren en de meer technische taalonderdelen als spelling en technisch lezen. Aan de andere kant is de correlatie tussen de toetsen Begrijpend luisteren en Begrijpend lezen (r = 0,51) lager dan verwacht. Maar zoals vermeld in hoofdstuk 2, staat bij begrijpend lezen het technisch verwerken van het schrift tussen de tekst en de lezer in en bij begrijpend luisteren speelt dit geen rol. Zeker aan het eind van groep 3 zal dit gegeven nog van grote invloed zijn, omdat nog veel leerlingen aan het eind van het 'technisch leesproces' staan. Bij begrijpend luisteren worden leerlingen niet gehinderd door 'het schrift' want in de toets Begrijpend luisteren wordt gebruikgemaakt van illustraties. In de toets wordt het begrip van de (gesproken) tekst gemeten zonder tussenkomst van geschreven tekst, die gecodeerd moet worden. We verwachten wel dat de correlatie tussen de toetsen Begrijpend luisteren en Begrijpend lezen hoger zal worden in de volgende jaargroepen, als het lezen meer geautomatiseerd zal verlopen.

6.2.3 Samenhang met de variabele leerjaar

In figuur 6.1 staan de vaardigheidsverdelingen voor de afnamemomenten M3 en met E3, waarvan de gemiddelden en varianties te vinden zijn in tabel 4.9 (overgenomen uit paragraaf 4.3.3).

Figuur 6.1 Vaardigheidsverdeling per afnamemoment



Tabel 4.9 Overzicht van de vaardigheidsverdelingen per afnamemoment (overgenomen uit paragraaf 4.3.3)

Afnamemoment	Aantal leerlingen	Gemiddelde vaardigheid	Standaardafwijking
M3	1198	47,63	8,72
E3	2037	50,08	8,68

Uit figuur 6.1 komt naar voren dat de vaardigheid Begrijpend luisteren groeit in de tijd. De verschillende toetsen op de onderscheiden meetmomenten zijn gesitueerd op dezelfde vaardigheidsschaal Begrijpend luisteren en laten een toename van de gemiddelde vaardigheid zien die verwacht mag worden op basis van leeftijd, ontwikkeling en de hoeveelheid genoten onderwijs.

6.2.4 Responsiviteit en stabiliteit

De toetsen in Cito Volgsysteem moeten veranderingen kunnen meten. Uit het kalibratieonderzoek is gebleken dat de opgaven op één onderliggende schaal Begrijpend luisteren liggen. De resultaten uit het normeringsonderzoek laten zien dat er verandering gemeten wordt, de gemiddelden per afnamemoment verschillen immers. Uit de (latente) correlaties in tabel 6.3 blijkt dat de correlatie hoog genoeg is om te kunnen beweren dat bijna alle leerlingen een zekere groei doormaken, maar niet zo hoog om te kunnen stellen dat dit voor alle leerlingen het geval is, dan wel dat de groei voor alle leerlingen even groot is.
In tabel 6.2 vinden we de aantallen leerlingen die op de beide normeringstijdstippen aan het onderzoek deelgenomen hebben. Bovenstaande is een onderbouwing voor het gegeven dat de toetsen Begrijpend luisteren voldoende responsief zijn om veranderingen te meten. Bovendien weerspiegelt de zeer hoge correlatie een grote stabiliteit in de uitkomsten van de toetsen. Daarbij moet men zich realiseren dat het op

elk afnamemoment om verschillende toetsen gaat die op dezelfde onderliggende vaardigheidsdimensie zijn geconstrueerd.
De hoge correlatie laat niet alleen zien dat de gemeten vaardigheid (begrijpend luisteren) zeer stabiel is, zij impliceert ook dat het goed gelukt is om de toetsen op dezelfde vaardigheidsschaal te situeren, hetgeen kan worden opgevat als een onderbouwing van de validiteit. De gevonden correlatie wijst ten slotte op hoge test-hertestbetrouwbaarheden per toets. Er is weliswaar geen test-hertestonderzoek uitgevoerd, maar wanneer een dergelijke hoge correlatie tussen *verschillende* toetsen op dezelfde vaardigheidsdimensie tussen twee meetmomenten wordt gevonden, mag men aannemen dat herhaalde afname van dezelfde toets met een korter afname-interval tot zeer hoge intercorrelaties zal leiden.

Tabel 6.2 Aantal gevolgde leerlingen op de verschillende normeringsmomenten

	M3	**E3**
M3	1192	
E3	710	2037

Tabel 6.3 Latente correlaties tussen leerlingen op de verschillende normeringsmomenten

	M3	**E3**
M3	1,000	
E3	0,854	1,000

6.2.5 Itemkarakteristieken

In deze paragraaf vatten we een aantal gegevens samen die betrekking hebben op de itemparameters van de toetsen Begrijpend luisteren groep 3. De gegevens worden gepresenteerd in tabel 6.4.
De gemiddelde moeilijkheidsgraad van de toetsen ligt op het door de toetsdeskundigen gewenste niveau, namelijk rond 0,75. De gemiddelde moeilijkheidsgraad voldoet daarmee aan het gestelde doel, namelijk een optimaal onderscheidend vermogen bij de groep met een lage of gemiddelde vaardigheid (zie verder hoofdstuk 5 over lokale meetnauwkeurigheid), terwijl de toetsen niet als moeilijk zullen worden ervaren door de doorsnee leerling. De moeilijkheidsgraad van de afzonderlijke opgaven kent een goede spreiding; er zijn zowel moeilijke als gemakkelijke opgaven in de toetsen opgenomen.
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; voor R_{it} is dat wel het geval.
De gemiddelde R_{it}-waarden zijn over het algemeen te kenschetsen als voldoende (0.20-0.29). In de tabel is te zien dat een aantal opgaven een lagere R_{it}-waarde dan .19 hebben: dit zijn opgaven met hoge p-waarden, dus zeer gemakkelijke opgaven. Voor de kwaliteit van de toets maakte het niet uit of deze opgaven in de selectie werden opgenomen of niet, maar we hebben besloten deze te handhaven, omdat we leerlingen van deze leeftijd (6-7 jaar) niet alleen wilden confronteren met relatief veel moeilijke opgaven, maar ook met een aantal eenvoudige opgaven.

Tabel 6.4 Samenvatting itemkenmerken voor de toetsen Begrijpend luisteren op de afnamemomenten M3 en E3

	M3			E3		
	P	**R_{it}**	**R_{ir}**	**P**	**R_{it}**	**R_{ir}**
gemiddeld	0,76	0,25	0,22	0,74	0,22	0,19
P10	0.56	0.20	0.19	0.59	0.19	0.18
Mediaan	0.80	0.23	0.22	0.82	0.23	0.21
P90	0.91	0.33	0.32	0.92	0.32	0.30
R_{it} < .19		3			7	

Zoals vermeld in paragraaf 4.3.2 is voor beide toetsen ook de constante ‘c’ berekend en komen er in beide toetsen alleen opgaven voor met een als ‘goed’ of ‘voldoende’ te kenschetsen c-waarde.

7 Samenvatting

In dit hoofdstuk wordt kort weergegeven wat in de voorafgaande hoofdstukken besproken is.
Nadat we in hoofdstuk 2 de uitgangspunten bij de toetsconstructie en in hoofdstuk 3 de inhoud van de toetsen uitvoerig hebben beschreven, hebben we in hoofdstuk 4 over het normeringsonderzoek gerapporteerd. We hebben daar verantwoord hoe het afnamedesign voor het kalibratieonderzoek was opgezet. Ook hebben we in hoofdstuk 4 aangegeven hoe we te werk zijn gegaan bij de steekproeftrekking. De wijze van steekproeftrekking en de controles achteraf (wat betreft percentage achterstandsleerlingen, geografische verdeling en mate van verstedelijking) wijzen uit dat de steekproef representatief genoemd kan worden voor de populatie van scholen in Nederland.
In hoofdstuk 5 rapporteerden we over de betrouwbaarheidscoëfficiënten. De betrouwbaarheidscoëfficiënten (MAcc's) zijn voor de toetsen Begrijpend luisteren voldoende. In figuur 5.1 is af te lezen hoe het is gesteld met de lokale meetnauwkeurigheid van de toetsen. Hierin is te zien dat de lokale meetnauwkeurigheid van de toetsen begrijpend luisteren M3 en E3 er redelijk uitziet. De standaardmeetfout loopt wat op met het stijgen van de vaardigheid, maar dit wijkt niet af van het beeld van andere toetsen.
Over validiteit rapporteerden we in hoofdstuk 6. In paragraaf 6.1 is ingegaan op de inhoudsvaliditeit van de toetsen, waarvoor de basis is gelegd in hoofdstuk 2 en 3. In paragraaf 6.2 is uitgebreid ingegaan op de begripsvaliditeit van de toetsen Begrijpend luisteren. Een belangrijke indicatie voor de validiteit van de opgaven uit de toetsen komt uit het kalibratieonderzoek (hoofdstuk 4). Daaruit is gebleken dat de opgavenverzameling waaruit de toetsen zijn samengesteld, beschreven kan worden met OPLM. Dat betekent dat de met de toetsen gemeten vaardigheid te verklaren is door een unidimensionaal model. Daarnaast kon worden vastgesteld dat de correlaties tussen de latente vaardigheden op twee opeenvolgende afnamemomenten hoog tot zeer hoog zijn. Dat betekent enerzijds dat de toetsen goede operationalisaties zijn van dezelfde onderlinge vaardigheid (i.e. begrijpend luisteren) en dat de stabiliteit van deze vaardigheid hoog is. Anders gezegd: als we weten wat de score van een leerling op een bepaald moment is, kunnen we daaruit goed afleiden wat zijn score op een volgend afnamemoment zal zijn. Ook is aangetoond dat de vaardigheidsscore gemiddeld toeneemt van afnamemoment tot afnamemoment, wat te verwachten is op basis van de toename in leeftijd en ontwikkeling en de vorderingen in het leerproces.
Een belangrijke aanwijzing voor convergente en discriminerende validiteit is af te leiden uit de correlaties tussen de toetsen Begrijpend luisteren met andere toetsen uit het Cito Volgsysteem primair onderwijs. Uit deze gegevens blijkt dat de scores op de toetsen Begrijpend luisteren sterker samenhangen met scores op meer semantische onderdelen, zoals woordenschat en begrijpend lezen, dan met scores op de meer 'technische' onderdelen, zoals rekenen-wiskunde, technisch lezen – leestechniek, technisch lezen - leestempo en spelling. Zo blijkt dat de samenhang van begrijpend luisteren met bijvoorbeeld woordenschat, een ondersteunende taalvaardigheid van begrijpend luisteren, met een correlatie van 0,77 aanzienlijk hoger correleert dan bijvoorbeeld de 'technische' taalvaardigheid spelling (r = 0,38). De gegevens over de itemkenmerken (moeilijkheidsgraad en item-totaalcorrelatie) laten tot slot een bevredigend beeld zien.

8 Literatuur

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

Berkel, S. van en N. Alberts (2009). *Woordenschat groep 3, Leerling- en onderwijsvolgsysteem primair onderwijs*. Arnhem: Cito.

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, M. Hilte en M. van der Zanden (2011). *Begrijpend luisteren groep 3.* Cito Volgsysteem primair onderwijs (LOVS). 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.

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*. Geheel herziene versie mei 2009, gewijzigde herdruk mei 2010. Amsterdam: NIP/COTAN

Expertgroep Doorlopende Leerlijnen (2008a). *Over de drempels met taal. De niveaus voor de taalvaardigheid.* Onderdeel van de eindrapportage van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen. Enschede: Expertgroep doorlopende leerlijnen TAAL EN REKENEN.

Expertgroep Doorlopende Leerlijnen (2008b). *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 (2009). *Een nadere beschouwing. Over de drempels met taal en rekenen.* Enschede: Expertgroep doorlopende leerlijnen TAAL EN REKENEN.

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

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.

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.

Greven, J., J. Letschert, SLO (2006). *Kerndoelen primair onderwij*s. 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.

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.

Janssen, J., F. Scheltens en J. Kramer (2006). *Rekenen-Wiskunde groep 3. Leerling- en onderwijsvolgsysteem primair onderwijs*. Arnhem: Cito.

Krom, R.S.H. (1992). *Luisteren 1*. 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. van Berkel en I. Jongen (2006). *Begrijpend lezen groep 3, Leerling- en onderwijsvolgsysteem primair onderwijs.* Arnhem: Cito.

Krom, R., I. Jongen en P. Roumans (2009), *Technisch lezen groep 3, Leestechniek en Leestempo, Leerling- en onderwijsvolgsysteem primair onderwijs*. Arnhem: Cito.

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).

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.

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

Ockey, G.J. (2007). Construct implications of including still image or video in computer-based listening tests. Language Testing, [Electronische versie], 24, 517-536.

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.

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., 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.

Weerden, J. van, K. Heesters, I. Jongen, F. van der Schoot, B. Hemker, N. Veldhuijzen en N. Verhelst (2006). *Balans van het spreekonderwijs op de basisschool 2. Uitkomsten van de peilingen in 2002 en 2003 halverwege en einde basisonderwijs en speciaal basisonderwijs.* Arnhem: Cito (PPON-reeks nr. 30).

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.

Wijs, A. de, S. van Berkel en R. Krom (2006). *Spelling groep 3. Leerling- en onderwijsvolgsysteem primair onderwijs*. Arnhem: Cito.

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 Voorbeelden van opgaventypen in de toetsen voor groep 3

Opgaven waarbij gegeven informatie in gesproken taaluitingen moet worden onthouden

- **Opgaventype 'combineren van inhoudselementen'**
  (Inspreekster:)
  Jim draagt veel te grote regenlaarzen en heeft een zwarte regenjas met witte stippen aan.
  Op welk plaatje zie je dat?


☐


☐


☐

- **Opgaventype 'selecteren van inhoudselementen'**
  (Inspreekster:)
  Jop is dol op sleeën, Milan op skiën en Jesse houdt erg van schaatsen.
  Op welk plaatje zie je Jop?

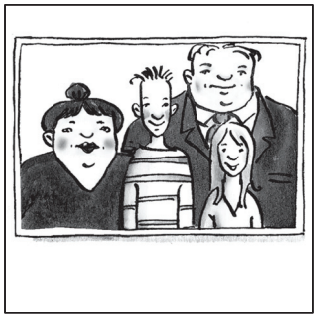

☐


☐

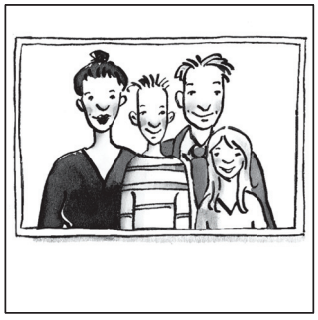

☐

Opgaven waarbij de samenhang binnen of tussen gesproken taaluitingen moet worden doorzien of waaruit geïmpliceerde informatie moet worden afgeleid

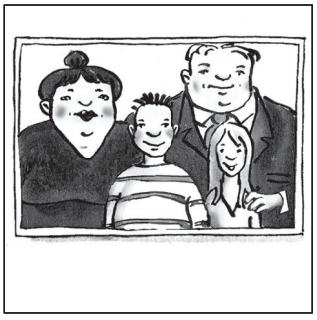
- **Opgaventype 'expliciete verbanden'**
  (Inspreekster:)
  Fleur, Vincent en hun ouders staan op de foto.
  Vincent is groter dan Fleur, maar kleiner dan zijn ouders.
  Vincent is ook steviger dan Fleur, maar minder stevig dan zijn ouders.
  Op welke foto staan Fleur en Vincent met hun ouders?



☐



☐



☐

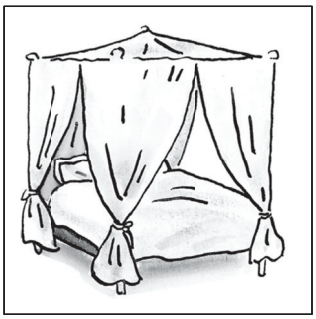
- **Opgaventype 'impliciete verbanden'**
  (Inspreekster:)
  'Ik zou ze eerst even strikken, anders struikel je nog', zegt Robbe tegen Sonia.
  Op welk plaatje zie je Sonia?

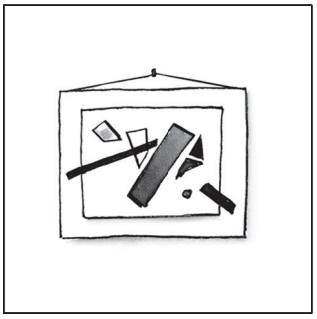
☐

☐

☐

- **Opgaventype ‘anaforische verwijsrelaties’**
  (Inspreekster:)
  In Senna’s slaapkamer hangt een modern schilderij.
  En tegen de muur staat een oud kastje.
  Senna vindt dat maar lelijk.
  Er staat ook een enorm bed.
  Wat vindt Senna lelijk?





☐ ☐ ☐

Opgaven waarbij de grammaticale organisatie van gesproken uitingen moet worden doorzien

- **Opgaventype ‘grammaticale constructies’**
  (Inspreekster:)
  Terwijl Niels gitaar en Ray piano speelt, zit Janniek achter het drumstel.
  Op welk plaatje zie je Niels?

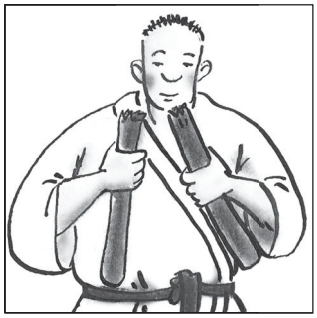


☐ ☐ ☐

Opgaven waarbij niet-letterlijk bedoeld taalgebruik in gesproken uitingen herkend moet worden

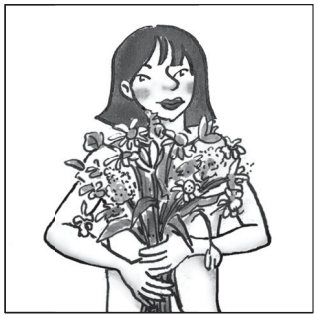
- **Opgaventype 'metaforisch taalgebruik'**
  (Inspreekster:)
  Na een dag hard werken, is Guust helemaal gebroken.
  Op welk plaatje zie je Guust?


☐


☐


☐

- **Opgaventype 'ironisch/sarcastisch taalgebruik'**
  (Inspreekster:)
  'Meid, je wordt bedánkt!', roept Aniek naar Isa.
  Op welk plaatje zie je Aniek?


☐

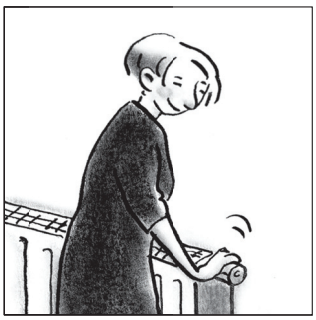

☐


☐

- **Opgaventype ‘verrichten van taalhandelingen’**
  (Inspreekster:)
  ‘Mama, het tócht hier’, zegt Fenna.
  Wat wil Fenna dat haar moeder doet?

☐

☐



☐

Opgaven waarbij geïmpliceerde informatie uit gesproken taaluitingen moet worden afgeleid

- **Opgaventype ‘kennis van het taalsysteem’**
  (Inspreekster:)
  ‘Nog een snufje zout en wat peper erbij. Dat maakt de smaak alleen nog maar voller’.
  Wie zegt dit?

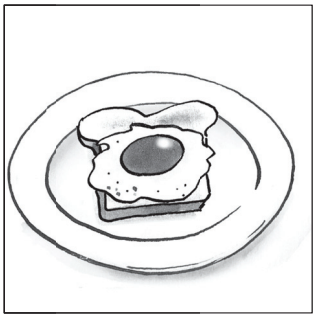
☐

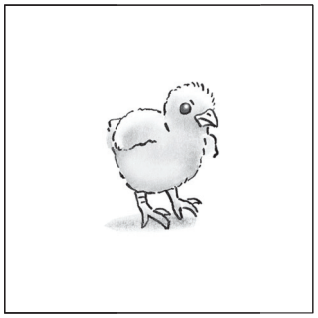
☐

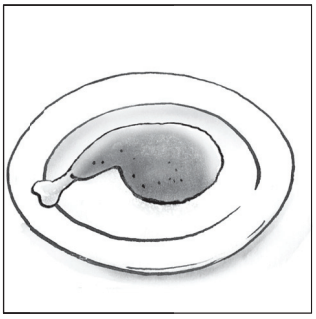
☐

eel 1

- **Opgaventype ‘kennis van gewoonten en gebruiken’**
  (Inspreekster:)
  Abel verzorgt samen met opa Koos de kippen.
  ‘Zo, wij hebben morgen weer een heerlijk ontbijtje’, zegt opa tegen Abel.
  Wát vindt Abel de volgende dag bij zijn ontbijt?

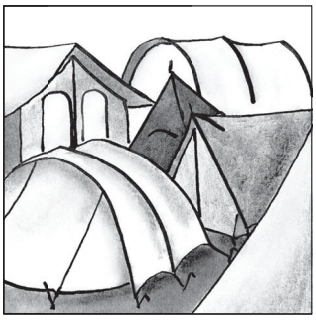

☐

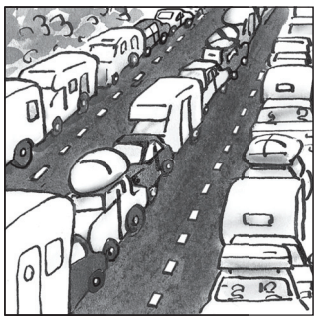

☐


☐

Opgaven waarbij de globale inhoud van gesproken taaluitingen moet worden herkend

- **Opgaventype ‘globale betekenistoekenning’**
  (Inspreekster:)
  Dit is het nieuws van 10 uur.
  De scholen in het zuiden van het land hebben vanaf vandaag vakantie.
  Er wordt daarom grote drukte op de Nederlandse wegen verwacht.
  U wordt gewaarschuwd voor vertraging naar uw vakantiebestemming.
  Welk plaatje past het best bij dit bericht?


☐


☐


☐

Cito maakt wereldwijd werk van goed en eerlijk toetsen en beoordelen. Met de meet- en volgmethoden van Cito krijgen mensen een objectief beeld van kennis, vaardigheden en competenties. Hierdoor zijn verantwoorde keuzes op het gebied van persoonlijke en professionele ontwikkeling mogelijk. Onze expertise zetten we niet alleen in voor ons eigen werk maar ook om advies, ondersteuning en onderzoek te bieden aan anderen.

Cito
Amsterdamseweg 13
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
F (026) 352 13 56
www.cito.nl

Klantenservice
T (026) 352 11 11
klantenservice@cito.nl

Fotografie: Ron Steemers