Wetenschappelijke verantwoording Cito Intelligentietest VO

Michel Hop en Herman van Boxtel

zeker weten

Wetenschappelijke Verantwoording
Cito Intelligentietest VO

Michel Hop
Herman van Boxtel

psychometrie Timo Bechger en Bas Hemker

Cito
Arnhem, 2013

© Cito B.V. Arnhem (2014)
Niets uit dit werk mag zonder voorafgaande schriftelijke toestemming van Cito B.V. worden openbaar gemaakt en/of verveelvoudigd door middel van druk, fotografie, scanning, computersoftware of andere elektronische verveelvoudiging of openbaarmaking, microfilm, geluidskopie, film- of videokopie of op welke wijze dan ook.

Inhoud

1 Inleiding

De Cito Intelligentietest VO maakt deel uit van het Cito Volgsysteem voortgezet onderwijs. Met de test meet men de algemene intelligentie van leerlingen aan de hand zes redeneertaken met opgaven in de categorieën Figuren, Woorden en Getallen. De resultaten kunnen helpen bij het bepalen van het meest geschikte onderwijsniveau voor leerlingen in de eerste drie leerjaren van het voortgezet onderwijs. Daarbij kan men de leerling vergelijken met andere leerlingen in hetzelfde leerjaar (op basis van leerjaar-georiënteerde normen). Ook de meer gebruikelijke leeftijdgeoriënteerde bepaling van het intelligentieniveau (IQ-bepaling door vergelijking met leeftijdgenoten) is mogelijk en wel voor leeftijdsgroepen van 11 tot en met 14 jaar.

De Cito-uitgave Intelligentietest VO bestaat – naast deze wetenschappelijke verantwoording – uit de volgende onderdelen:

- Handleiding
- Toetsboekjes
- Antwoordformulier
- Ouderfolder.

Daarnaast zijn nog beschikbaar:

- Scoringsservice, met als resultaat een papieren rapportage per leerling
- Online rapportagesysteem (RAVAS) met daarin leerlingrapportages en groepsoverzichten.

In deze wetenschappelijke verantwoording wordt verslag gedaan van de constructie van de test. Daarbij wordt aandacht besteed aan het theoretisch kader van waaruit de test is opgezet. Het constructieproces en de afname van de proefversie worden gerapporteerd, evenals de wijze waarop de normering van de test heeft plaatsgevonden. Daarnaast worden alle analyses en resultaten besproken die het de gebruiker mogelijk maken om een uitspraak te doen over de belangrijkste kenmerken, waaronder de normering, de betrouwbaarheid en validiteit van dit instrument.
Tot slot is een hoofdstuk over scoring en interpretatie van de test toegevoegd. Daarin is niet alleen alle informatie opgenomen die in de handleiding voor de docent is vermeld ten behoeve van de interpretatie, maar wordt ook ingegaan op de resultaten die in deze verantwoording zijn gerapporteerd voor zover deze voor de interpretatie relevant zijn.
Deze verantwoording is vooral bedoeld voor gebruikers en andere professionals die zich een beeld willen vormen van de kwaliteit van de test. Tezamen met het hierbovengenoemde testmateriaal levert deze wetenschappelijke verantwoording alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit op de volgende aspecten:

- Uitgangspunten voor de testconstructie
- De kwaliteit van het testmateriaal
- De kwaliteit van de handleiding
- Normen
- Betrouwbaarheid
- Begripsvaliditeit
- Criteriumvaliditeit

Daarnaast is de verantwoording bestemd voor docenten die zich nader willen verdiepen in de achtergronden van de test en voor gedragswetenschappers die de afname van de test begeleiden.
Ook personen en instanties die in tweede instantie testscores van leerlingen in handen krijgen, vinden in deze verantwoording voldoende aanknopingspunten voor de interpretatie van deze scores.

2 Gebruiksdoel en uitgangspunten voor de testconstructie

De Intelligentietest Cito Volgsysteem voortgezet onderwijs (in het vervolg van deze wetenschappelijke verantwoording simpelweg aangeduid als Intelligentietest VO) kan op verschillende manieren en momenten worden ingezet. In paragraaf 2.1 bespreken we eerst de meetpretentie van de test door antwoord te geven op de vraag wat we met dit instrument beogen te meten. We staan kort stil bij de context waarin de test vooral kan worden gebruikt: het geven van advies over het best passende onderwijsniveau. Deze context stelt eisen aan de samenstelling van de testbatterij en de wijze van normeren. In paragraaf 2.2 gaan we vervolgens concreter in op de gebruiksdoelen en functies van de test. We doen dit wat uitgebreider dan in de handleiding voor de afname (Cito, 2012) om de relatie tussen de meetpretentie en de gebruiksmogelijkheden van de test te verhelderen. Tevens hopen we daarmee duidelijk te kunnen maken welke validiteitsaanspraken we maken. In paragraaf 2.3 gaan we in op de theoretische achtergronden die aan de basis liggen van deze test.

2.1 Meetpretentie en karakterisering van de test

Algemene intelligentie
Intelligentie is geen eenduidig omschreven begrip waarover in de wetenschappelijke literatuur een duidelijke consensus bestaat. Wie op zoek gaat naar een definitie, treft dan ook een groot scala aan omschrijvingen aan. De termen ‘intelligentie’ en ‘algemene intelligentie’ worden daarbij op een weinig consistente wijze door elkaar gebruikt. Met de aanduiding ‘algemene intelligentie’ wil men soms tot uitdrukking brengen dat het om een samenstel van vermogens gaat, soms ook dat er sprake lijkt te zijn van een grote hoeveelheid gemeenschappelijke variantie in scores op verschillende deeltaken. Deze wordt dan geïnterpreteerd in termen van een psychologisch of psychometrisch construct dat ook wel aangeduid wordt met het begrip g-factor.
Met de termen intelligentie en algemene intelligentie (het onderscheid vinden we niet erg relevant) doelen we op:

> “het geheel van cognitieve en verstandelijke vermogens dat nodig is om kennis te verwerven, daar op een goede wijze gebruik van te maken, teneinde problemen op te kunnen lossen die een vast omschreven doel en structuur hebben.”
> (Resing en Drenth, 2001, p. 19).

In deze definitie komt naar voren dat intelligentie zowel van belang is in relatie tot probleemoplossingsvaardigheid als in relatie tot kennisverwerving. Intelligente kinderen zijn beter in het oplossen van een groot scala van welomschreven problemen, maar zijn ook beter in het verwerven van nieuwe kennis en het toepassen daarvan. Sternberg (1985) maakt in dit opzicht een belangrijk onderscheid tussen inzicht (‘insight’) in de aard en oplosbaarheid van een probleem enerzijds en de opslag en automatisering van kennis en probleemoplossingsstrategieën anderzijds. Intelligente leerlingen begrijpen dus sneller dan minder intelligente leerlingen hoe een probleem in elkaar zit en komen gemakkelijker tot een oplossing, ze leren ook sneller en efficiënter, slaan de kennis adequater op en weten deze ook beter beschikbaar te maken. Dit laatste is een andere manier om te zeggen dat intelligente leerlingen over het algemeen betere schoolprestaties laten zien: zij behalen hogere scores op leervorderingentoetsen.

In bovenstaande definitie hebben we algemene Intelligentie omschreven als een geheel, een optelsom van cognitieve en verstandelijke vermogens. Daarbij sluiten we aan bij een lange traditie in het meten van intelligentie. Zo heeft Wechsler, als ontwerper van waarschijnlijk de wereldwijd meest gebruikte reeks van algemene intelligentietests, in de handleiding en verantwoording bij de Amerikaanse versie van de WISC-R, intelligentie omschreven als: “the overall capacity of an individual to understand and cope with the world around him. This definition conceives of intelligence as an overall or global entity; that is, a multi-determined and multi-faceted entity rather than an independent, uniquely-defined trait. It avoids singling out any ability, however esteemed (e.g. abstract reasoning), as crucial or overwhelmingly important.”’ (Wechsler, 1974 in

Geelhoed, Struiksma & Moesker (2009), p. 384). Elders drukt Wechsler zich in vergelijkbare bewoordingen uit, waar hij intelligentie benoemt als: “the aggregate or global capacity of the individual (to act purposefully, to think rationally and to deal effectively with his environment)” (1944, p. 3). In deze benadering is het gebruikelijk om algemene intelligentie te meten door een testbatterij af te nemen die bestaat uit een aantal verschillende subtests en taken. Welke taken dat precies zijn hangt mede af van de keuze van een specifieker theoretisch kader en van de functie die de testbatterij moet vervullen. We komen hier nog uitgebreider op terug in paragraaf 2.3 over de theoretische achtergronden van de test.

De context: advisering over het best passende onderwijsniveau: intelligentie versus leervorderingen
Als het erom gaat de leerling, zijn ouders of verzorgers, docenten of docententeams en gedrags-wetenschappers van advies te dienen over het best passende onderwijsniveau, wordt in de regel een beroep gedaan op twee soorten instrumenten die ieder op hun eigen manier iets zeggen over de cognitieve vermogens van de leerling.
Op de eerste plaats kan men proberen om zo precies mogelijk in kaart te brengen wat de huidige schoolprestaties van de leerling zijn. Men gebruikt dan, naast de reguliere gegevens die over een leerling bekend zijn (zoals cijfers voor schoolwerk, overhoringen en proefwerken; rapportcijfers), leervorderingen-toetsen zoals de toetsen van het Cito Volgsysteem voortgezet onderwijs. Men probeert zo op grond van leervorderingen die de leerling tot op dat moment heeft gemaakt een antwoord te geven op de vraag of de leerling in het huidige onderwijs op zijn plek zit of wat zijn kansen op succes zijn in toekomstig onderwijs. Goede leervorderingentoetsen zeggen veel over die succeskansen. Voor een deel komt dit omdat dit type instrumenten indirect een afgewogen meting vormen van een aantal eigenschappen die van groot belang zijn voor toekomstig schoolsucces, zoals leertempo, concentratie, motivatie en doorzettingsvermogen. Daarnaast doet zo’n toets – en ook het onderwijs waarvan de opbrengst bij de leerling door middel van de toets zichtbaar wordt gemaakt – indirect een beroep op de intelligentie van het kind.
Het voordeel van de leervorderingentoets is echter tegelijkertijd ook zijn nadeel, want op deze manier zorgen andere eigenschappen van de leerling ervoor dat men er niet zeker van is dat de leerling via de toets laat zien waartoe hij cognitief gezien werkelijk in staat is. Men brengt in kaart “wat er uit komt”, waarbij men er niet zeker van is of dat overeenkomt met “wat er in zit”. Niet alle leerlingen presteren immers conform hun capaciteiten. Door bijvoorbeeld een gebrek aan motivatie, concentratie en inzet steken sommige leerlingen minder op van het onderwijs dan op basis van hun capaciteiten zou mogen worden verwacht en minder dan wat hun even intelligente leeftijdgenoten die wél gemotiveerd zijn om te leren van het onderwijs opsteken. En daarmee komt als alternatief de intelligentietest in beeld, die beoogt op een zo zuiver mogelijke manier in beeld te brengen waartoe de leerling in cognitief opzicht in staat is.
Een intelligentietest drukt deze cognitieve mogelijkheden in de regel uit in een getal dat intelligentiequotiënt (IQ) wordt genoemd. Daarbij dient men zich wél te realiseren dat *hét* intelligentiequotiënt niet bestaat. De uitkomst zal immers voor een deel bepaald worden door de precieze aard en samenstelling van het instrument. Wij komen daar zo dadelijk nog op terug.
Tegen deze achtergrond kan de Cito Intelligentietest VO op verschillende manieren, in verschillende situaties, en ter beantwoording van verschillende onderzoeksvragen worden ingezet. Daarbij gaat het er steeds vooral om welke plaats in het onderwijs *op dit moment* het beste past bij de leerling. Cito heeft met de test niet de ambitie ook het uiteindelijke schoolsucces (haalt de leerling een diploma in een bepaald onderwijstype, hoe lang doet hij daarover) exact te kunnen voorspellen.
In de volgende sectie gaan we verder in op de verschillende doelen en functies van de test, maar voor een beter begrip willen we eerst enkele belangrijke eigenschappen van het instrument bespreken.

Taken in de intelligentietest
Om de cognitieve en verstandelijke vermogens van leerlingen (intelligentie) goed in kaart te brengen is een evenwichtige keuze van taken noodzakelijk. Dat geldt voor elke intelligentietest. Zoals gezegd, *hét* IQ bestaat immers niet. Elke intelligentietest is gebaseerd op zijn eigen specifieke visie op wat intelligentie is en op de wetenschappelijke onderbouwing daarvan. Cito maakt geen fundamenteel onderscheid tussen intelligentietests en leervorderingentoetsen, althans niet ten aanzien van de taken die in beide soorten instrumenten zijn ondergebracht. Er zijn dus geen wezenlijke verschillen zoals die wél bestaan tussen bijvoorbeeld een motoriektest en een beroepeninteressetest. Intelligentietests en leervorderingentoetsen worden beide opgevat als samenhangende hulpmiddelen tot onderzoek van de intellectuele prestaties. Er is

in die taken slechts sprake van een gradueel, door onze onderwijscultuur bepaald verschil. Het gaat in beide gevallen om taken waarbij de leerling zijn cognitieve vermogens moet inzetten. Sommige van die taken sluiten meer aan bij op school systematisch onderwezen technieken en inhouden (leervorderingentoetsen). Andere taken richten zich meer op intellectuele functies die in het onderwijs minder of minder rechtstreeks aan bod komen (intelligentietests). De taken zijn met andere woorden te plaatsen op een continuüm, waarvan de uitersten duidelijk zijn te onderscheiden, maar waarin geen nauwkeurige grens te trekken is (zie van Boxtel, Snijders & Welten, 1982). Bij de Cito Intelligentietest VO zijn de taken zo gekozen dat zij zo dicht mogelijk tegen het uiteinde van dit continuüm liggen. Dat wil zeggen: deze taken doen een beroep op basale redeneervaardigheden en hogere mentale processen die zo min mogelijk het resultaat zijn van kennis en (leer)ervaring. Op die manier kunnen eventuele discrepanties tussen intelligentie en leervorderingen, dus tussen “wat er in zit” en “wat er uit komt”, zo scherp mogelijk in beeld worden gebracht.

Normering

Historisch gezien was het IQ (intelligentiequotiënt) een echt quotiënt in die zin dat de zogeheten mentale leeftijd werd gedeeld door de kalenderleeftijd en vervolgens met 100 vermenigvuldigd. Aan deze werkwijze kleven belangrijke technische bezwaren. Wereldwijd worden IQ’s daarom tegenwoordig opgevat als deviatie-IQ, waarbij de leerling wordt vergeleken met een groep leerlingen die qua leeftijd ongeveer even oud zijn en waarbij wordt aangenomen dat intelligentie in de populatie normaal verdeeld is. In het deviatie-IQ komt tot uitdrukking in welke mate de testscore van een leerling overeenkomt, dan wel afwijkt van het gemiddelde van zijn of haar leeftijdsgroep. Voor elke leeftijdsgroep wordt de gemiddelde score vastgesteld op 100, met een standaarddeviatie van 15.

Intelligentietests worden meestal dus naar leeftijd genormeerd, zo ook de Cito Intelligentietest VO. Er zijn afzonderlijke normen per leeftijdsgroep met een breedte van één jaar. Dat is vooral nodig omdat intelligentie bij kinderen en jeugdigen nog volop in ontwikkeling is, wat grotendeels is toe te schrijven aan rijping van de hersenen. Omdat de vraagstellingen waarbij de Cito Intelligentietest VO kan worden ingezet zich vooral afspelen in de onderbouw van het voortgezet onderwijs, is de test genormeerd voor de leeftijden die deze fase ‘afdekken’, dus voor 11- tot 14-jarigen. Daarbij is zorgvuldig rekening gehouden met de verdeling van leerlingen in verschillende schooltypen en –niveaus, en is ook onderzoek gedaan in het regulier basisonderwijs en in vormen van speciaal onderwijs.

Sommige intelligentietests, en dan gaat het vooral om intelligentietests die in en ten behoeve van het onderwijs worden gebruikt (vergelijk bijvoorbeeld de NIO, de Nederlandse Intelligentietest voor Onderwijs-niveau (van Dijk & Tellegen, 2004)), zijn niet naar kalenderleeftijd, maar naar leerjaar genormeerd. Dat wil zeggen dat de leerling niet vergeleken wordt met zijn exacte leeftijdgenoten (bijvoorbeeld de 12-jarigen), maar met leerlingen die in hetzelfde leerjaarcohort zitten, bijvoorbeeld alle leerlingen in leerjaar 1, respectievelijk leerjaar 2 enzovoorts. Men doet dit door op deze manier aan te sluiten bij de natuurlijke context waarmee school en docenten zich geconfronteerd zien. Leerlingen zitten immers in groepen bijeen die gekenmerkt worden door de gebruikelijke mix van kalenderleeftijden. Als het gaat om het plaatsen van leerlingen in verschillende typen of niveaus van vervolgonderwijs (bijvoorbeeld havo of vwo in vervolg op een tweejarige brugklas havo/vwo) is een dergelijke vergelijking ook handiger en meer op zijn plaats. Omdat het bij de Cito Intelligentietest VO voor een belangrijk deel om vragen gaat die betrekking hebben op plaatsing van leerlingen, is ervoor gekozen om naast het gebruikelijke leeftijd-IQ ook een leerjaar-IQ te berekenen. Beide manieren om de cognitieve vermogens van een leerling uit te drukken zullen bij de meeste leerlingen leiden tot waarden die met elkaar overeenkomen of zeer dicht in elkaars buurt liggen. Alleen voor de leerlingen die duidelijk ouder dan wel jonger zijn dan het gemiddelde voor hun leerjaargroep zal er sprake zijn van enig verschil.

Er is nog een tweede reden om ook een leerjaar-IQ te berekenen. In volgsystemen voor het voortgezet onderwijs (zoals het Cito Volgsysteem voortgezet onderwijs), worden (toets-)prestaties van leerlingen genormeerd naar leerjaar en niet naar leeftijd. Als men dus van leerlingen hun intelligentie en leer-vorderingen onderling wil vergelijken, ligt het voor de hand om voor beide een leerjaarnormering te hanteren.

In dit kader is het zinnig om nog eens extra te benadrukken dat cognitieve vermogens deels afhankelijk zijn van ontwikkelingen die zich in het brein voltrekken. Ook al is intelligentie tot op zekere hoogte genetisch

bepaald, dit wil niet zeggen dat daarmee iemands cognitieve vermogens volledig vastliggen en voortdurende hetzelfde blijven. Genen bepalen weliswaar mede tot welke cognitieve prestaties iemand gedurende zijn levensloop in staat is en bij metingen van die prestaties op verschillende leeftijden zal sprake zijn van een zekere stabiliteit. Wanneer iemand in zijn cognitieve prestaties steeds met leeftijdgenoten wordt vergeleken (in de vorm van een deviatie-IQ), zal deze vergelijking meestal laten zien dat de resultaten van de verschillende metingen niet heel erg ver uit elkaar lopen. Maar tegelijkertijd kan niet worden ontkend dat er groei en ontwikkeling plaatsvindt. De gemiddelde tweejarige is niet tot dezelfde cognitieve prestaties in staat als de gemiddelde twaalfjarige. Elk kind en elke jongere doorloopt een ontwikkelingstraject, maar de timing daarvan is bij iedereen anders. Bij sommigen maken bepaalde delen van de hersenen op leeftijd x een snelle ontwikkeling door die bij anderen (nog) achterwege blijft. Ook al past men nog zulke goede normen toe en ook al is de test of toets die men gebruikt nog zo betrouwbaar, met deze exacte individuele ontwikkelingstrajecten kan het instrument geen rekening houden. Concreet kan dit betekenen dat bijvoorbeeld de ene 12-jarige op het moment van testafname al net wél zo'n specifieke ontwikkeling in het brein heeft doorgemaakt (een ontwikkeling die hem in zijn cognitieve vermogens in het voordeel stelt in vergelijking met een deel van zijn leeftijdgenoten), terwijl de andere 12-jarige die ontwikkeling nog moet doormaken en dus relatief in het nadeel is (Ramsden, Richardson, Josse, Thomas, Ellis, Shakeshaft, Seghier, & Price, 2011). Of, met andere woorden, timingverschillen in de cognitieve ontwikkeling gaan ten koste van de stabiliteit en de voorspelbaarheid over tijd. Het kan dus erg zinnig zijn om onderzoek naar de cognitieve vermogens van leerlingen te herhalen, zeker op momenten dat zij opnieuw voor een keuze staan (bijvoorbeeld doorstroom na de brugklas) en / of bij leerlingen van wie de schoolprestaties ernstig mee- of tegenvallen in vergelijking met wat men op basis van eerder onderzoek verwachtte. Juist om deze reden hebben uitkomsten van intelligentiemetingen bij kinderen en tieners maar een beperkte houdbaarheid van ten hoogste één à twee jaar.

Verdeling naar schooltype en –niveau
Sommige vraagstellingen veronderstellen dat bekend is hoe de verdeling van IQ-gegevens binnen bepaalde schooltypen en –niveaus eruit ziet. Daarom wordt in het leerlingrapport het leerjaar-IQ van de leerling afgezet tegen het gemiddelde leerjaar-IQ van elk afzonderlijk schoolniveau. In de rapportage worden de volgende schoolniveaus onderscheiden: vmbo bb met leerwegondersteunend onderwijs, vmbo bb, vmbo kb, vmbo gt , havo en vwo.

Collectieve en individuele afname
De Cito Intelligentietest kan collectief (klassikaal) en individueel worden afgenomen, op voorwaarde dat er maatregelen worden genomen die verhinderen dat er kan worden afgekeken of samengewerkt. Het zal van het type vraag waarop men antwoord hoopt te vinden afhangen voor welke vorm van afnemen men een voorkeur heeft. Informatie over doel en functie van de test is te vinden in de volgende paragraaf.

2.2 Doel en functie van de test

Manieren van gebruik
Men kan de test op verschillende manieren gebruiken. Daarbij staat steeds de vraag centraal wat op basis van cognitieve capaciteiten het meest geschikte onderwijsniveau is voor een leerling binnen het voortgezet onderwijs.

Als u alle leerlingen van een klas of leerjaar wilt onderzoeken op eventuele verschillen tussen capaciteiten en leervorderingen neemt u de test bij al deze leerlingen af. U gebruikt de test dan signalerend. De test kan ook voor individuele leerlingen worden ingezet. Dat doet u bijvoorbeeld als u een vermoeden over een leerling, bijvoorbeeld over onderpresteren, nader wilt onderbouwen.

De uitslag van de Intelligentietest is ondersteunend bij de vraag of een leerling geplaatst is in het onderwijsniveau dat het best bij de leerling past en in situaties waarin u beslissingen moet nemen over het onderwijsniveau waarin een leerling het meest tot zijn recht zal komen. Het leerlingrapport geeft informatie over de gemiddelde intelligentie en de spreiding daarvan per onderwijsniveau. Door de score van een leerling met

deze informatie te vergelijken krijgt u inzicht in welk niveau het best bij hem of haar past op basis van zijn of haar intelligentie.

Ten slotte kunt u de Intelligentietest VO gebruiken bij de verwijzing naar lwoo en pro. Cito gaat ervan uit dat de Intelligentietest VO wordt opgenomen in de lijst van hiervoor toegestane instrumenten.

De resultaten van de Intelligentietest geven samen met Toets 0 t/m 3 van het Cito Volgsysteem voortgezet onderwijs zinvolle informatie over wat u van een leerling mag verwachten. Door informatie te verzamelen over intelligentie én leervorderingen kan elke leerling in zijn leerproces optimaal begeleid worden en wordt de beslissing over het vervolgonderwijs uitgebreid ondersteund.

Vereiste deskundigheid
De Cito Intelligentietest VO is een eenvoudig af te nemen instrument om de intelligentie vast te stellen. De test kan rechtstreeks door scholen en docenten worden afgenomen, er is geen gedragswetenschapper voor nodig zolang de onderzoeksvraag betrekking heeft op plaatsing van de leerling binnen het voortgezet onderwijs. De test kan worden afgenomen door één van de docenten bij wie de leerling in de klas zit, bijvoorbeeld de klassendocent of de mentor. Dat heeft grote voordelen omdat de leerlingen doorgaans hun docenten goed kennen en met hen vertrouwd zijn. Bovendien sluit deze manier van afnemen goed aan bij de Toets 0 t/m 3 van het Cito Volgsysteem voortgezet onderwijs. Voorwaarde is dat de docent zich goed voorbereidt op zijn of haar rol als afnameleider. De afname-instructies zijn daarom uitgebreid beschreven in de handleiding.
Het is ook mogelijk om iemand anders dan een docent (bijvoorbeeld een gedragswetenschapper) als afnameleider van de test in te zetten. In deze situatie gelden uiteraard dezelfde voorbereidingseisen.
Ook kan de school ervoor kiezen zich bij de afname van de test te laten begeleiden door bijvoorbeeld een schoolbegeleidingsdienst en de afname te laten uitvoeren door een gekwalificeerd testleider onder supervisie.
Interpretatie van de resultaten kan plaatsvinden door een docent of in het docententeam, vooropgesteld dat het om onderzoeksvragen gaat die zich beperken tot de plaatsing van de leerling in de onderwijsniveaus binnen het voortgezet onderwijs. Bij andersoortige vragen dient een gedragswetenschapper of begeleidingsdienst te worden ingeschakeld.

Doelgroep
De doelgroep van de Intelligentietest zijn alle leerlingen in leerjaar 1 tot en met 3 van het reguliere voortgezet onderwijs. Voor deze leerlingen kan met de test een leerjaar-IQ worden vastgesteld. Van de leerlingen van 11 tot en met 14 jaar wordt er daarnaast een op leeftijd gebaseerde intelligentiescore vastgesteld. Voor de leeftijdnormering zijn gegevens verzameld in het voortgezet- en basisonderwijs.
Het onderzoek in het basisonderwijs (groep 6, 7 en 8) was nodig om de normgroepen naar leeftijd representatief te maken. Er zijn ook afnames verricht in het (voortgezet) speciaal onderwijs en het speciaal basisonderwijs, zodanig dat het leeftijd-IQ gebaseerd is op een dwarsdoorsnede uit de gehele populatie. Dit betekent echter niet dat de test ook geschikt zou zijn voor leerlingen uit deze speciale vormen van onderwijs. Bij de schaling van de testopgaven is ervan uitgegaan dat deze leerlingen anders reageren op de opgaven dan andere leerlingen. Zij zijn daarom niet opgenomen in de datasets waarop kalibratieanalyses zijn uitgevoerd.

Voor bepaalde groepen leerlingen kan afname van de Intelligentietest VO problemen opleveren. Te denken valt aan leerlingen met NLD, slechtziende en blinde leerlingen. In het algemeen gaat het om leerlingen bij wie de Eindtoets Basisonderwijs alleen in aangepaste versie kon worden afgenomen (bijvoorbeeld een gesproken, vergrote of brailleversie). Van de Intelligentietest VO zijn geen aangepaste versies beschikbaar en daarom kan de test bij deze leerlingen niet worden afgenomen.

De test is wel bruikbaar voor dyslectische leerlingen. Uit het normeringsonderzoek is gebleken dat dyslectische leerlingen niet anders scoren dan niet-dyslectische leerlingen (voor meer informatie hierover zie hoofdstuk 6). Dyslectische leerlingen kunnen de test onder dezelfde condities maken als niet-dyslectische leerlingen.

De Intelligentietest VO bevat een belangrijke verbale component. Dat betekent dat bijvoorbeeld een beroep wordt gedaan op de kennis van verbale begrippen en het hanteren van relaties tussen verbale begrippen. Hiervoor is gekozen omdat deze verbale component van belang wordt geacht voor de mogelijkheden die de leerling heeft om van onderwijs te kunnen profiteren en een verdere schoolloopbaan in het voortgezet onderwijs tot een succes te maken. Dit zou kunnen betekenen dat leerlingen die het Nederlands niet als moedertaal hebben of nog te kort in Nederland verblijven om het Nederlands afdoende te beheersen, bij deze test moeilijkheden ondervinden. Uit analyses is gebleken dat kinderen die thuis een andere taal spreken dan het Nederlands gemiddeld lager scoren op deze intelligentietest dan kinderen die thuis Nederlands spreken (zie hoofdstuk 6). Houdt u daar rekening mee bij de beslissing over het afnemen van de test en de interpretatie van de resultaten.

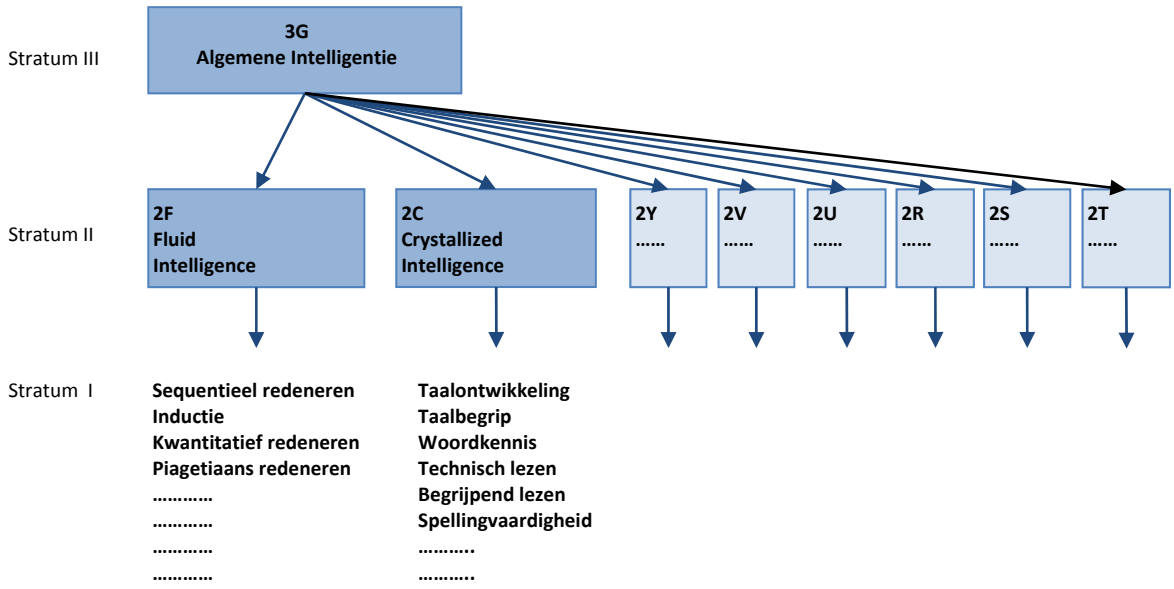
2.3 Theoretische uitgangspunten

Intelligentie en leervorderingen: één continuüm?
Eerder hebben we al aangegeven dat er veel verschillende opvattingen zijn over wat intelligentie nu precies is. Daarbij hebben we gekozen voor een bepaalde definitie: algemene Intelligentie als een geheel, een optelsom van cognitieve en verstandelijke vermogens. Bovendien hebben we benadrukt dat de taken die gekozen worden in intelligentietests enerzijds niet wezenlijk verschillen van de taken in een leervorderingentoets (beide soorten taken doen een beroep op die cognitieve en verstandelijke vermogens), maar anderzijds toch een specifiek karakter dragen. Taken in intelligentietests hebben met elkaar gemeen dat zij over het algemeen zo min mogelijk een beroep doen op de kennis en leerervaring die via het formele onderwijs op school wordt opgedaan.
De opvatting dat taken in leervorderingentoetsen en intelligentietests niet fundamenteel van elkaar verschillen is gebaseerd op literatuur over taken in verschillende soorten instrumenten die bedoeld zijn om het cognitieve presteren te onderzoeken (zie van Boxtel et al., 1982). Uit deze literatuur bleek, dat correlaties binnen de groep van leervorderingentoetsen en binnen de groep van intelligentietests van dezelfde grootte-orde zijn als tussen taken die uit beide groepen afkomstig zijn. Men kan taken onderscheiden die vrij direct aansluiten bij de op school systematisch onderwezen kennisinhouden en vaardigheden. Andere taken richten zich vooral op het onderzoek van de intellectuele functies, taken die in het onderwijs minder of niet rechtstreeks aan bod komen, zoals het strikt logisch redeneren, het analyseren en manipuleren van ruimtelijke structuren. De eerste taken zijn meer kenmerkend voor leervorderingentoetsen, de andere meer kenmerkend voor intelligentietests. Sommige taken zoals het benoemen van synoniemen en tegenstellingen vindt men in beide soorten instrumenten terug. In van Boxtel et al. (1982) worden analyses gerapporteerd op data die verzameld werden met de ISI-Reeks Vorm III en worden de uitkomsten besproken van analyses met betrekking tot eerdere versies deze testreeks waarin zowel leervorderingen- als intelligentiesubtests zijn opgenomen. De resultaten van deze analyses bleken ondersteunend voor de opvatting dat er tussen beide soorten tests alleen een gradueel, door onze onderwijscultuut bepaald verschil bestaat. Beide testsoorten zijn onder te brengen op een continuüm, waarvan de uitersten duidelijk zijn te onderscheiden, maar waarin geen nauwkeurige grenslijn te trekken is.

Factoranalytische benadering en keuze van subtests
De keuze van taken in een intelligentietest is niet onbelangrijk. Deze taken bepalen namelijk welke aspecten van de algemene intelligentie worden benadrukt en in hoeverre deze (samen) een goede afspiegeling vormen van de algemene intelligentie.
Tegelijkertijd is het vooral belangrijk om taken te kiezen die een goede criteriumvaliditeit ten opzichte van schoolprestaties en leervorderingen hebben zonder dat zij daarmee samenvallen. Taken zoals die zijn opgenomen in de NIO, de voormalige ISI-Reeks of de Cito Intelligentietest Eindtoets Basisonderwijs laten correlaties zien van boven 0,70. In dit opzicht is het opmerkelijk dat de veel gebruikte WISC-III (ondanks subtests als Vocabulaire en Rekenvaardigheid) zich tevreden lijkt te stellen met correlaties die niet meer dan 25% verklaarde variantie impliceren. Zo worden in de herziene handleiding en verantwoording van deze test (Kort, Schittekatte, Dekker, Verhaeghe, Compaan, Bosmans, & Vermeir, 2005) correlaties met schoolcijfers van tussen 0,37 en 0,53 gerapporteerd. De WISC-IV (die in Nederland niet zal worden

uitgebracht (zie de website van de uitgever)) belooft in dit opzicht veel meer; vergelijk bijvoorbeeld Flanagan en Kaufman (2004).
Aan de andere kant is het de bedoeling om een zo groot mogelijk contrast te creëren tussen leervorderingen enerzijds en taken die zoveel mogelijk het potentieel van de leerling representeren anderzijds, waarbij (op school) aangeleerde kennis als middel om die taken tot een goed einde te brengen zoveel mogelijk wordt vermeden.
Voor het selecteren van taken (subtests) in intelligentietests is vooral het werk van Carroll (1993) belangrijk geweest. Carroll heeft aan de hand van gegevens die verzameld zijn met een groot aantal verschillende intelligentietests ontdekt, dat het mogelijk is op verschillende niveaus van algemeenheid naar de scores op intelligentietests te kijken en ook welke dimensies van intelligentie op elk niveau een rol spelen. Zijn werk is gebaseerd op een factor-analytische insteek met factoranalyse als het belangrijkste methodologische werktuig. Die niveaus van algemeenheid noemt hij “strata” (lagen). Carroll onderscheidt drie strata, waarbij Stratum III het hoogste niveau van algemeenheid representeert en Stratum I het meest specifieke.
Een vereenvoudigde afbeelding van dit drie-strata-model is terug te vinden in Figuur 2.1.

Figuur 2.1 Vereenvoudigde weergave van het Carroll-structuurmodel van cognitieve vaardigheden



In Stratum III is er sprake van één algemene factor (algemene intelligentie; ‘general intelligence’). Sommigen spreken ook wel over de G-factor. Op de interpretatie van deze eerste, algemene factor komen we zo dadelijk nog wat uitgebreider terug.
Ten aanzien van stratum II onderscheidt Carroll acht tweede orde factoren die hij als volgt benoemt (we handhaven de oorspronkelijke Engelstalige labels, zie p. 626):

- Fluid Intelligence (2F)
- Crystallized Intelligence (2C)
- General Memory and Learning (2Y)
- Broad Visual Perception (2V)
- Broad Auditory Perception (2U)
- Broad Retrieval Ability (2R)
- Broad Cognitive Speediness (2S)
- Processing Speed (Reaction Time Decision Speed) (2T)

Van deze tweede-orde factoren zijn de zogenoemde ‘Fluid Intelligence’ (2F) en ‘Crystallized Intelligence’ (2C) (deze termen zijn wat lastig in het Nederlands te vertalen) de twee belangrijkste omdat deze aspecten samen een groot deel van de algemene intelligentie (G) bepalen. Ze zijn in de literatuur ook het beste gedocumenteerd.

Elke dimensie op niveau II kan op niveau I nog verder worden gespecificeerd. Zo noemt Carroll voor ‘Fluid Intelligence’ de aspecten ‘General sequential reasoning’, ‘Induction’ en ‘Quantitative reasoning’ (algemeen sequentieel redeneren, inductie en kwantitatief redeneren). Voor ‘Crystallized Intelligence’ komen aspecten als taalontwikkeling, taalbegrip, vocabulaire, begrijpend lezen, spellingvaardigheid, fonetisch coderen en gevoeligheid voor grammatica naar voren. Zoals te zien is, hebben deze laatste aspecten direct of indirect veel te maken met taal en zijn ze voor een flink deel het resultaat van leren en ervaring. Ze komen op die manier van alle algemene cognitieve vaardigheden het meest in de buurt van de onderwerpen en vaardigheden die op school worden onderwezen. Hoewel deze vaardigheden wel degelijk deel uitmaken van de algemene intelligentie (althans in sommige opvattingen daarvan) is ervoor gekozen om in de selectie van taken voor de Cito Intelligentietest VO de nadruk te leggen op basale redeneervaardigheden en de hogere mentale processen zoals deze met name naar voren komen als aspecten van ‘Fluid Intelligence’; vaardigheden en processen dus die veel minder het resultaat zijn van kennisverwerving en (leer)ervaring. De reden hiervoor is om zo scherp mogelijk onderscheid te kunnen maken tussen intelligentie als cognitief potentieel (“wat zit erin”) enerzijds en leerprestaties (als realisatie van dat potentieel, “wat komt eruit?”) anderzijds. Daar komt nog bij dat de factor “Fluid Intelligence” het sterkst van alle tweede orde factoren de algemene intelligentie bepaalt. Leereffecten spelen bij deze factor een relatief bescheiden rol.
De keuze van de taken voor de Cito Intelligentietest VO wordt daarnaast sterk bepaald door overwegingen met betrekking tot inhoud. In de problemen waarvoor we ons in onze cultuur gesteld zien en in het onderwijs dat voorbereidt op het (cognitief) functioneren in deze cultuur spelen (ruimtelijke) figuren, woorden en getallen een belangrijke rol. Om die reden is gekozen voor een opzet met opgaven waarin geredeneerd moet worden met figuren, woorden en getallen. De opgaven zijn zó gekozen dat steeds het redeneren voorop staat en dat de meer “crystallized” aspecten van intelligentie naar de achtergrond zijn gebracht. Dus geen woordkennis, symboolkennis of cijferen en de kennis van de tafels, maar opgaven die betrekking hebben op de relatie tussen respectievelijk (eenvoudige, hoogfrequente) woorden, figuren en getallen.

Kanttekeningen bij de gekozen benadering
Als uitkomst van bovengenoemde overwegingen en toepassing van de genoemde selectiecriteria zijn in de Cito Intelligentietest VO zes verschillende redeneertaken ondergebracht. Bij twee taken wordt geredeneerd met figuren (het domein ‘Figuren’), bij twee taken wordt geredeneerd met woorden (het domein ‘Woorden’) en bij twee taken wordt geredeneerd met getallen (het domein ‘Getallen’). Kortom, de Cito intelligentietest VO is hiermee een test met een brede spreiding van taken op een vaardigheid die het sterkst samenhangt met algemene intelligentie en zo min mogelijk invloed ondervindt van leren en ervaring.
Bij deze keuzes zijn kanttekeningen te plaatsen. Men kan zich afvragen of men bij de interpretatie van een somscore over de onderscheiden deeltaken nog wel van intelligentie, dan wel algemene intelligentie, c.g. G-factor kan spreken. Op de eerste plaats past hier een relativering van het construct G zelf, een relativering die met name is ontleend aan een artikel van van der Maas, Dolan, Grasman, Wicherts, Huizenga, & Raijmakers (2006). Van der Maas et al. (2006) wijzen op het onderscheid tussen G als een *psychometrisch* en als *psychologisch* construct. Zij gaan daarbij uit van het verschijnsel dat scores op cognitieve taken (zoals dat het geval is bij de subtests van een intelligentietest of –batterij) vrijwel zonder uitzondering een patroon van positieve intercorrelaties laten zien dat uitnodigt tot de interpretatie dat er sprake is van een onderliggende latente variabele die de score op de onderscheiden taken mede beïnvloedt. Bij factoranalyse op zo’n correlatiematrix (aangeduid met de term ‘positive manifold’) komt in de regel een dominante eerste orde factor naar voren met een dominante eerste eigenwaarde die op te vatten is als een samenvattende psychometrische maat of index voor de (sterkte van de samenhang binnen de) ‘positive manifold’. De auteurs plaatsen vervolgens een vraagteken bij de interpretatie van deze index als een psychologisch construct, dat wil zeggen als een fundamentele cognitieve factor of de mogelijke oorsprong daarvan. Het zou heel goed kunnen dat een ander mechanisme het bekende intercorrelatie-patroon verklaart. De auteurs bieden een nieuwe, alternatieve verklaring, in termen van een ‘mutualism model’ dat kan worden opgevat als een ontwikkelingspsychologisch, dynamisch procesmodel: “In this explanation, we assume that in the intial phase of development, cognitive processes are uncorrelated. During development, the positive manifold emerges as a consequence of mutually beneficial interactions between these processes.” (p. 855). Door middel van simulatiestudies op basis van een logistisch

groeimodel met interactiecomponent weten de auteurs overtuigend aannemelijk te maken dat over verloop van tijd een positief intercorrelatiepatroon ('positive manifold') ontstaat op basis van de assumptie "that these cognitive processes have mutual beneficial or facilitating relations[1]." (p. 845). Met andere woorden, het is helemaal niet nodig om de positieve intercorrelaties tussen subtests in een intelligentietestbatterij die de operationalisatie vormen van verschillende cognitieve processen te 'verklaren' door het postuleren van een algemene intelligentiefactor waarvan het psychologische bestaan lastig te bewijzen valt.
Op de tweede plaats kunnen we stellen dat de redeneerfactor waarop we de nadruk hebben gelegd in de keuze van taken voor deze test, zelf een relatief brede en fundamentele factor is. Breed in die zin dat redeneervaardigheden het hart van de intelligente vermogens vormen en van toepassing zijn op verschillende problemen en probleemsituaties (zoals bijvoorbeeld blijkt uit de keuze van taken op verschillende domeinen, figuren, getallen, woorden). Breed ook in die zin dat de redeneerfactor als tweede orde factor de sterkste samenhang met de 'G- factor' laat zien. Fundamenteel, in zoverre dat 'fluid intelligence' waarschijnlijk vooraf gaat aan 'crystallized intelligence'. Van der Maas et al. (2006) die aantonen dat hun model (zie boven) ook kwesties als de hiërarchie in factoren zoals Carroll die aanbrengt en het onderscheid tussen 'fluid' en 'crystallized inteligence' kan verklaren, zeggen hierover: "Cattell's (1971) distinction between fluid intelligence (*gf*) en crystallized intelligence (*gc*) can be accomodated in the mutualism model. Crystallized intelligence is thought to develop by the interaction of fluid intelligence and cultural experience". Dit citaat sluit verder ook goed aan bij onze poging om in de keuze van taken een zo groot mogelijk contrast te creëren tussen intelligentie ("wat zit erin") en leervorderingen ("wat komt eruit"). Of, anders gezegd, taken die de 'crystallized intelligence'-factor representeren (zoals de in intelligentietests veel voorkomende taken als woordenschat, synoniemen, tegenstellingen en woordcategorieën) liggen op het eerder besproken continuüm ergens in het overgangsgebied tussen de basale redeneertaken en de leervorderingentoetsen in (zie bijvoorbeeld analyses met betrekking tot de taken in de ISI-Reeks zoals gerapporteerd in van Boxtel et al., 1982). Het zijn juist dit soort taken die we zoveel mogelijk hebben vermeden.

Opbouw en samenstelling van de test
In het voorafgaande hebben we aangegeven dat de Cito Intelligentietest VO stamt uit de klassiek psychometrische traditie. Van Dijk en Tellegen (2004) omschrijven deze als volgt: "De theoretische achtergrond is gefundeerd in de factoranalyse en met name in de theorie van de hiërarchische opbouw van de intelligentie waarbij naast een algemene intelligentiefactor ook groepsfactoren worden onderscheiden" (p. 5). Qua opzet en doelstelling is de test in grote lijnen goed vergelijkbaar met tests die ook uit deze traditie voortkomen, zoals de Cito Intelligentietest Eindtoets Basisonderwijs, de NIO en de GIVO. Verschil is dat het accent op 'fluid'-redeneertaken nog groter is dan bij de genoemde tests.

De test bestaat uit twee boekjes. Elke boekje bestaat uit 55 opgaven verdeeld over drie domeinen, te weten Figuren, Woorden en Getallen.

In beide testboekjes komen deze zelfde drie domeinen voor, met dien verstande dat elk domein twee varianten kent en in elk testboekje van elk domein maar één variant voorkomt.
De varianten zijn:
- Figuren: classificatie en matrix
- Woorden: classificatie en analogieën
- Getallen: reeksen en analogieën

[1] In principe is het mogelijk dat ook zonder deze veronderstelde wederzijds ondersteunende of faciliterende processen een 'positive manifold' ontstaat, zie Van der Maas et al. (2006).

De domeinen en soorten opgaven zijn als volgt over de opgavenboekjes verdeeld.

Opgavenboekje A		**Opgavenboekje B**	
Soort opgave	*Aantal*	*Soort opgave*	*Aantal*
Figuurclassificatie	15	Figuurmatrix	15
Woordclassificatie	15	Woordanalogie	15
Getalreeks	15	Getalanalogie	15
Woordclassificatie	5	Woordanalogie	5
Figuurclassificatie	5	Figuurmatrix	5
Totaal aantal opgaven	55	Totaal aantal opgaven	55

Op deze manier zijn er voldoende opgaven om de betrouwbaarheid per domein te garanderen en is er voldoende variatie in de opgaven om de aandacht vast te houden. De opgaven in elk boekje zijn per domein (in grote lijnen) geordend naar moeilijkheid. De vijf moeilijkste opgaven van Woorden en Figuren zijn achteraan in de boekjes geplaatst om te voorkomen dat in geval van tijdnood de opgaven van het domein Getallen niet gemaakt kunnen worden en zouden ontbreken in de leerlingrapportage.

In het volgende hoofdstuk gaan we nader in op de meetpretentie en taakbeschrijving van elk van deze deeltaken.

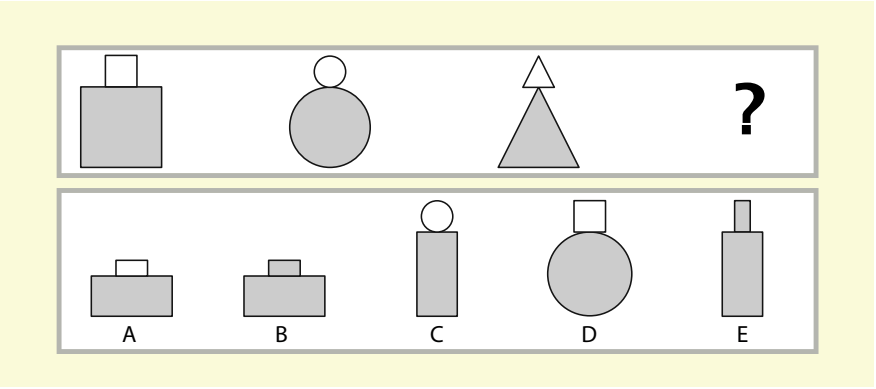
3 Constructie van de test

In dit hoofdstuk geven we eerst een beschrijving van de afzonderlijke taken (paragraaf 3.1). We beschrijven per taak steeds de meetpretentie en de taak die de leerling moet uitvoeren om tot een correcte oplossing van de opgaven te komen.[2] Vervolgens beschrijven we in paragraaf 3.2 de wijze waarop de test is geconstrueerd. We besteden aandacht aan de itemconstructie, het uittesten van de items in een pilot-onderzoek en de uiteindelijke selectie van de items. De definitieve versie van de test is gekalibreerd op basis van een IRT-meetmodel. De beschrijving daarvan, de gehanteerde procedures en informatie over de passing van het meetmodel zijn te vinden in hoofdstuk 4.

3.1 De samenstelling van de Intelligentietest VO

Figuurclassificatie
In *Figuurclassificatie* wordt de inductieve redeneervaardigheid gemeten waarbij deze wordt toegepast op figuren. De taak is inductief omdat de leerling eerst voor de in het bovenste vak gegeven figuren een algemene regel moet afleiden, namelijk welke kenmerken deze gemeenschappelijk hebben. Vervolgens moet deze algemene regel worden toegepast door een figuur uit het onderste vak te selecteren met dezelfde kenmerken.

Een voorbeeld:



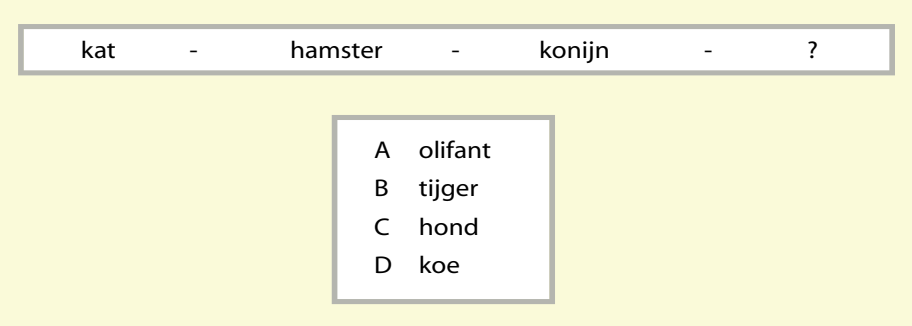
De vraag bij deze opgave is:
Welk figuur uit het onderste vak moet op de plaats van het vraagteken staan?

Oplossing
Het correcte antwoord is A. Want de algemene regel die moet worden afgeleid uit het bovenste vak houdt in dat een correcte figuur samengesteld is uit twee dezelfde vormen boven elkaar, waarbij de onderste steeds de grotere is met de kleur grijs en de bovenste wit. Figuur A is de enige figuur uit het onderste vak die aan deze regel voldoet.

Woordclassificatie
In Woord*classificatie* wordt inductieve redeneervaardigheid gemeten door deze toe te passen op woorden. De taak is inductief omdat de leerling eerst voor de in het bovenste vak gegeven woorden een algemene regel moet afleiden, namelijk welke kenmerken deze gemeenschappelijk hebben. Vervolgens moet deze algemene regel worden toegepast door een woord uit het onderste vak te selecteren met dezelfde kenmerken.

[2] De hier opgenomen voorbeelden zijn ontleend aan de voorbeeldopgaven in de test.

Een voorbeeld:

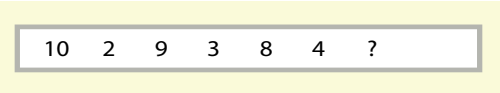


De vraag bij deze opgave is:
Welk woord uit het onderste vak moet op de plaats van het vraagteken staan?

Oplossing
Het correcte antwoord is C. Want de algemene regel die moet worden afgeleid uit het bovenste vak houdt in dat alle woorden betrekking hebben op dieren die als huisdier worden gehouden. Het antwoord C is het enige antwoord uit het onderste vak dat aan deze regel voldoet.

Getalreeks
In *Getalreeks* wordt de kwantitiatieve, inductieve redeneervaardigheid gemeten. De taak is kwantitatief omdat er gewerkt wordt met getallen waarbij de relatie tussen de getallen wiskundig beschreven kan worden. De kwantitatieve redeneervaardigheid kan zowel inductieve als deductieve redeneerprocessen betreffen. In dit geval wordt de inductieve redeneervaardigheid gemeten omdat uit de gegeven reeks van getallen een algemene regel moet worden afgeleid die de relatie tussen de getallen weergeeft. Vervolgens moet deze regel worden toegepast om het volgende getal in de reeks te bepalen.

Een voorbeeld:

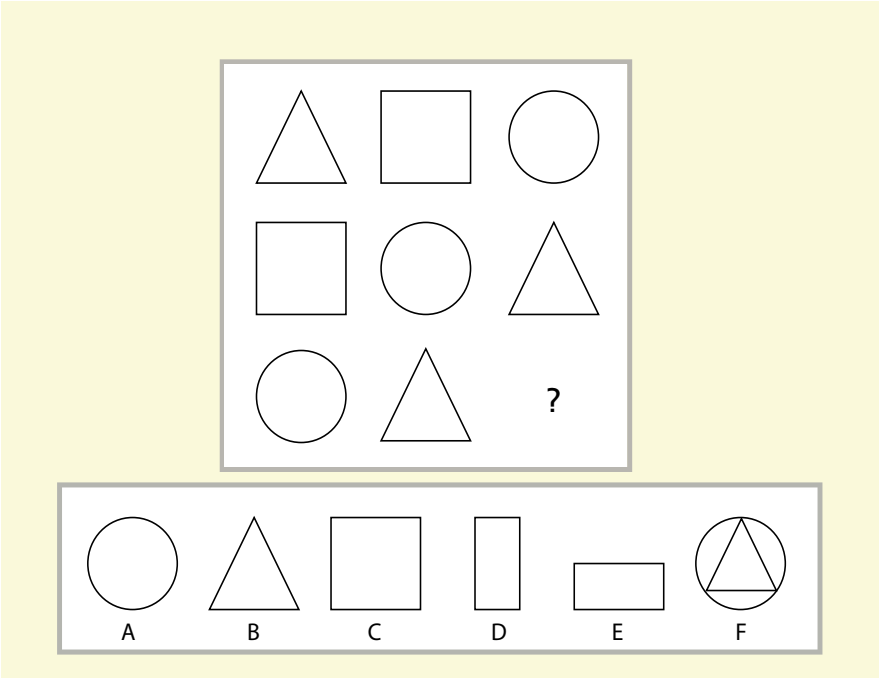


De vraag bij deze opgave is:
Welk getal moet op de plaats van het vraagteken komen te staan?

Oplossing
Het goede antwoord is 7. Want de algemene regel die de relatie russen de getallen beschrijft houdt in dat er, uitgaande van het eerste getal in de reeks, steeds alternerend een getal wordt opgeteld dan wel afgetrokken waarbij het op te tellen of af te trekken getal steeds met 1 afneemt. Dus in de eerste stap wordt 8 afgetrokken van het eerste getal, vervolgens 7 opgeteld bij het resulterende tweede getal, dan min 6, plus 5, min 4 en ten slotte plus 3 om het getal op de plaats van het vraagteken te genereren.

Figuurmatrix
In *Figuurmatrix* wordt inductieve redeneervaardigheid gemeten door deze toe te passen op figuren. De taak is inductief omdat de leerling in de matrix een algemene regel moet afleiden die de relatie tussen de figuren beschrijft, zowel horizontaal, verticaal als diagonaal. Vervolgens moet deze algemene regel worden toegepast door een figuur uit het onderste vak te selecteren dat overeenkomt met de afgeleide regel en dus past in de matrix.

Een voorbeeld:

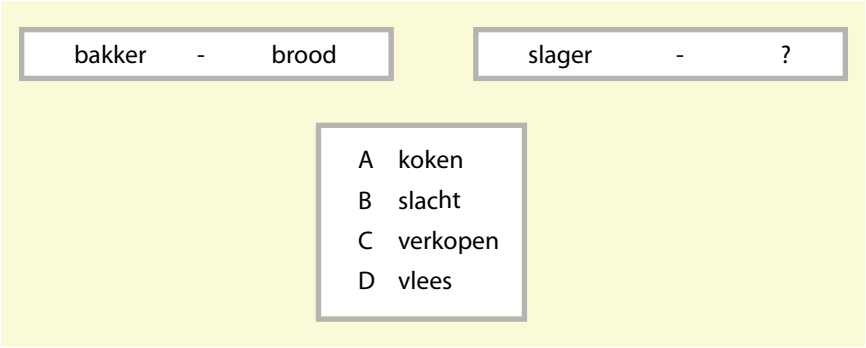


De vraag bij deze opgave is:
Welk figuur uit het onderste vak moet op de plaats van het vraagteken staan?

Oplossing
Het correcte antwoord is C. Want de algemene regel die moet worden afgeleid uit de martix is dat in elke rij (horizontaal) en kolom (verticaal) steeds een driehoek, vierkant en cirkel staat. De matrix wordt correct ingevuld wanneer Figuur C wordt geselecteerd want alleen dan gaat de algemene regel op.

Woordanalogie
In *Woordanalogie* wordt de inductieve redeneervaardigheid gemeten waarbij deze wordt toegepast op woorden. De taak is inductief omdat de leerling eerst voor de in het bovenste linkervak gegeven woorden een algemene regel moet afleiden die de relatie tussen deze twee woorden beschrijft. Vervolgens moet deze regel worden toegepast op de woorden in het bovenste rechtervak door een woord uit het onderste vak te selecteren.

Een voorbeeld.

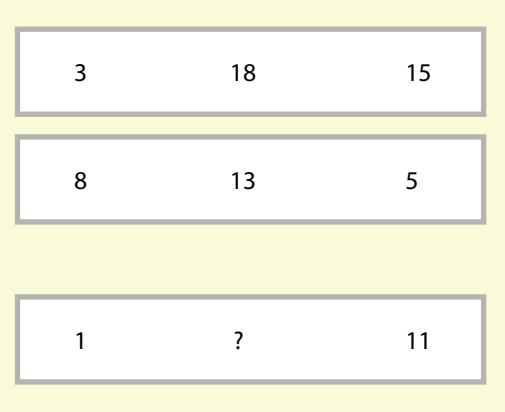


De vraag bij deze opgave is:
Welk woord hoort er bij ‘slager’, zoals ‘brood’ bij ‘bakker’ hoort?

Oplossing
Het correcte antwoord is D. Want de algemene regel die moet worden afgeleid uit het bovenste linkervak houdt in dat het eerste woord steeds de winkelier of verkoper aanduidt en het tweede woord betrekking heeft op het product dat deze verkoopt. Het antwoord D is het enige antwoord uit het onderste vak dat aan deze regel voldoet: de ‘slager’ verkoopt ‘vlees’ zoals de ‘bakker’ ‘brood’ verkoopt.

Getalanalogie
In *Getalanalogie* wordt de kwantitiatieve, inductieve redeneervaardigheid gemeten. De taak is kwantitatief omdat er gewerkt wordt met getallen waarbij de relatie tussen de getallen wiskundig beschreven kan worden. De kwantitatieve redeneervaardigheid kan zowel inductieve als deductieve redeneerprocessen betreffen. In dit geval wordt inductieve redeneervaardigheid gemeten omdat uit de twee gegeven reeksen van getallen een algemene regel moet worden afgeleid die de relatie tussen de getallen voor elke reeks op dezelfde manier weergeeft. Vervolgens moet deze regel worden toegepast op de onderste reeks om het ontbrekende (gevraagde) getal te bepalen.

Een voorbeeld:



De vraag bij deze opgave is:
Welk getal moet op de plaats van het vraagteken komen te staan?

Oplossing
Het goede antwoord is 12. Want de algemene regel die de relatie russen de getallen in de horizontale reeksen beschrijft houdt in dat het getal in het midden van de reeks steeds de som is van het getal links en het getal rechts in die reeks. De som van 1 en 11 is gelijk aan 12.

3.2 Constructie en selectie van testitems

Het constructieproces
De opgaven voor de intelligentietest werden ontwikkeld in twee groepen die bestonden uit respectievelijk drie en vier constructeurs met een afgeronde universitaire opleiding psychologie die bekend waren met de inhoud van (intelligentie)tests en ervaring hadden in de afname en toepassing daarvan. De leden van de constructiegroepen werden door een toetsdeskundige van Cito opgeleid in het construeren van items.
Zij kregen vervolgens in een aantal ronden steeds opdracht om elk twaalf opgaven van een bepaald type te construeren. In een vergadering werden de opgaven vervolgens kritisch besproken en indien nodig aangepast. Ongeschikte opgaven werden afgekeurd. Er hebben in totaal zeven vergaderingen plaats-gevonden.
Vooraf werd een aantal criteria voor de constructie vastgelegd en met de constructeurs besproken. Hieronder beperken we ons tot de belangrijkste criteria.
Voor de opgaven met woorden diende men woorden te kiezen met een hoge woordfrequentie. Het gaat bij alle taken immers om vaardigheden in het ontdekken van regels met betrekking tot de relatie tussen

woorden en niet om woordkennis. Items met te moeilijke woorden of waarin erg specifieke kennis noodzakelijk was voor het goed kunnen bepalen van de relatie tussen de woorden vielen af.
Voor de opgaven met getallen was het van belang de reeksen voldoende lengte te geven zodat er voldoende aanknopingspunten waren om de te ontdekken regel te kunnen afleiden. Ook hier gaat het er vooral om de relatie tussen de getallen te ontdekken en niet de rekenvaardigheid van de leerlingen te meten. Daarom is ervoor gekozen om getallen op te nemen waarmee gemakkelijk gerekend kan worden; voor de bewerking ervans is weinig rekenwerk vereist.
Voor de opgaven met figuren was het belangrijkste criterium dat de figuren voldoende duidelijk waren. Er is gekozen voor het gebruik van abstracte en geometrische figuren. Figuren waar mensen een voorkeur of sterk gevoel bij zouden kunnen hebben zijn zoveel mogelijk vermeden. Daarom zijn, op een enkele uitzondering na, geen afbeeldingen van mensen, dieren en dingen gebruikt.

Pilots
Op deze manier zijn in eerste instantie 355 items geconstrueerd over (in dat stadium nog) zeven categorieën. In juni 2008 zijn deze opgaven voor het eerst uitgetest op basis van een onvolledig design bij 289 leerlingen van vier scholen voor voortgezet onderwijs. Onvolledig wil hier zeggen dat de leerlingen steeds slechts een gedeelte van de opgaven kregen voorgelegd. Deze proeftoets had als belangrijkste doel om een eerste selectie te kunnen maken uit de opgaven. Opgaven die om verschillende redenen onbruikbaar bleken vielen af. In november 2008 is een tweede proeftoets uitgevoerd met als doel psychometrische kenmerken van de items te verzamelen. In totaal zijn 280 items die uit de eerste ronde waren overgebleven geproeftoetst en hebben 1452 leerlingen van zes scholen de toetsen gemaakt.
De scholen zijn willekeurig geselecteerd. Ongeveer twintig scholen zijn telefonisch benaderd met de vraag of zij wilden deelnemen aan een proefonderzoek voor de ontwikkeling van een intelligentietest. De zes genoemde scholen zegden hun medewerking toe. De scholen waren gelegen in Gelderland (3), Zeeland (1), Flevoland (1) en Overijssel (1). De groep scholen en kinderen is niet landelijk representatief omdat dit voor constructiedoeleinden (in tegenstelling tot normeringsdoeleinden) niet nodig is. Er is uiteraard wel gestreefd naar een redelijk evenwichtige verdeling over onderwijsniveaus, leerjaren, leeftijden en geslacht om de bruikbaarheid en kwaliteit van de opgaven voor alle verschillende categorieën van leerlingen te kunnen uitproberen. De verdeling van de leerlingen naar leeftijd, leerjaar, onderwijsniveau en sekse is gegeven in tabel 3.1.

Tabel 3.1 Verdeling van leerlingen in het proefonderzoek naar leeftijd, leerjaar, onderwijsniveau en sekse

Kenmerk	*Aantal leerlingen*	*Percentage*
Leeftijd		
12 jaar	176	12,2
13 jaar	536	37,1
14 jaar	486	33,6
15 jaar	217	15,0
16 jaar	30	2,1
onbekend	7	-
Leerjaar		
leerjaar 1	553	38,1
leerjaar 2	593	40,8
leerjaar 3	306	21,1
Onderwijsniveau		
vmbo	532	36,6
vmbo-havo	254	17,5
havo	221	15,2
[Tabel 3.1 vervolg]		
havo-vwo	226	15,6
vwo	199	13,7
onbekend	20	-
Sekse		
Jongen	696	48,2
Meisje	749	52,8
onbekend	7	-

De 280 items waren vrijwel gelijk verdeeld over de zeven categorieën zoals blijkt uit onderstaand overzicht.

FA	FC	FM	WA	WC	GA	GR
40	39	40	40	40	41	40

FA = Figuuranalogie
FC = Figuurclassificatie
FM = Figuurmatrix
WA = Woordanalogie
WC = Woordclassificatie
GA = Getalanalogie
GR = Getalreeks

De items werden volgens een onvolledig design verdeeld over acht verschillende taakboekjes (‘booklets’); waarbij elk boekje zeventig opgaven bevatte, uit elke categorie tien. In elk boekje was de helft van het aantal onderzochte opgaven gelijk aan de opgaven van een ander boekje. Op deze manier konden alle opgaven met elkaar verbonden worden.
De afnamematerialen zijn naar de scholen gestuurd met een uitgebreide afname-instructie. De scholen dienden zelf docenten aan te stellen als afnameleider. De afname kende een centrale klassikale instructie door de docent waarna de leerlingen geheel zelfstandig de test dienden te maken. Leerlingen namen zelf per categorie de instructie in het taakboekje met de bijbehorende voorbeeldopgaven door.
De acht verschillende boekjes werden door ongeveer gelijke aantallen leerlingen gemaakt, gemiddeld door 186,5 leerlingen. Hieruit valt af te leiden dat voor elke opgave van gemiddeld 373 leerlingengegevens beschikbaar kwamen. De gemiddelde moeilijkheidsgraad per taak was 0,67, de gemiddelde score 46,5 met een standaarddeviatie van 11,17. Deze samenvattende gegevens hebben betrekking op alle onderzochte

opgaven en zeggen dus nog weinig over de kenmerken van de uiteindelijk geselecteerde opgaven. Deze worden verderop in dit hoofdstuk in detail gerapporteerd.

Selectie van items en samenstelling van de definitieve testversie

Na de proeftoetsing in november 2008 is de definitieve test samengesteld onder gebruikmaking van de bekende itemparameters uit de klassieke testtheorie. Daarbij werden de gegevens gebruikt van de onderzoeksgroep die eerder in tabel 3.1 is beschreven.

Allereerst is gekozen voor een evenwichtige verdeling van domeinen. Om die reden is het aantal figuur-domeinen teruggebracht van drie naar twee. Besloten werd om de categorie Figuuranalogie te laten vallen. Deze redeneertaak doet ook een beroep op ruimtelijke oriëntatie, waarbij redeneervaardigheden relatief minder dominant zijn dan bij de andere 2 figuurtaken. De uiteindelijke testversie bevat dus steeds twee taken per domein (figuren, woorden en getallen).

Voor de zes resterende deeltaken werden de aantallen items geselecteerd die bij de opzet van het instrument waren voorzien (zie hoofdstuk 2). Dat wil zeggen steeds twintig items voor de taken Figuurclassificatie, Figuurmatrix, Woordclassificatie en Woordanalogie en vijftien items voor de taken Getalreeks en Getalanalogie. Bij de selectie van de definitieve items is rekening gehouden met een aantal criteria. Er is gestreefd naar een zo breed mogelijke inhoudelijke dekking van het domein en een zo groot mogelijke variatie in de aard van de items per domein. Daarnaast speelde de psychometrische kwaliteit van de items een belangrijke rol, zoals deze naar voren kwam in de moeilijkheidsgraad en itemresttotaal-correlatie. De p-waarde diende tussen de 0,30 en 0,90 te liggen en per leerjaar minimaal 0,20 en maximaal 0,90 te bedragen. De r_{ir}-waarde diende bij voorkeur boven de 0,20 te liggen. Voor een bepaalde deeltaak diende daarnaast sprake te zijn van een goede opbouw en verdeling naar moeilijkheidsgraad. Per deeltaak werd gestreefd naar een gemiddelde moeilijkheidsgraad voor de drie leerjaren tesamen van omstreeks 0,70.

In tabel 3.2 zijn de kenmerken van de deeltaken en items samengevat. In plaats van de r_{ir}-waarden die bij de itemselectie werden gebruikt vermelden we in de tabel de r_{it}-waarden omdat voor r_{it} kwaliteitscriteria bekend zijn en voor r_{ir} niet (zie Evers, Lucassen, Meijer & Sijtsma, 2010).

*Tabel 3.2 Samenvatting van itemkenmerken (moeilijkheidsgraad en r_{it}) per deeltaak (waarden * 100)*

Deeltaak	*p-waarde*			*r_{it}-waarde*		
	laagste	*hoogste*	*gemiddeld*	*laagste*	*hoogste*	*gemiddeld*
Figuurclassificatie	42	89	70,6	17	55	37,9
Figuurmatrix	49	89	74,8	28	54	43,1
Woordclassificatie	49	95	81,6	22	48	34,1
Woordanalogie	39	95	71,0	17	56	35,5
Getalreeks	35	89	71,6	42	62	50,9
Getalanalogie	50	84	70,0	38	64	55,0

Uit de tabel blijkt dat voor de meeste deeltaken de gemiddelde moeilijkheidsgraad, zoals beoogd, in de buurt ligt van de 0,70. Alleen de gemiddelde moeilijkheidsgraad van Figuurmatrix ligt met 0,75 iets hoger dan de beoogde waarde en die van Woordclassificatie met 0,82 duidelijk hoger. Niettemin is het gemiddelde discriminerende vermogen van laatstgenoemde deeltaak met 0,34 goed te noemen, al is het wat lager dan dat van de andere deeltaken. De gemiddelde r_{it}-waarden van de deeltaken Getalreeks en Getalsanalogie zijn hoog te noemen en duiden op een homogeen karakter. De r_{it}-waarden voor de individuele items liggen slechts in twee van de 110 gevallen (1,8%) onder de grens van 0,20 die door de COTAN wordt aangemerkt als de ondergrens voor ‘voldoende’ (beide met een waarde van 0,17). Volgens diezelfde COTAN-criteria zijn negen items (8,2%) als voldoende te kwalificeren en 99 (90%) als goed (met r_{it}-waarden van ≥ 0,30).

We kunnen concluderen dat we erin geslaagd zijn om voor de verschillende deeltaken opgaven te selecteren die voldoen aan de beoogde doelen. De hier aangegeven indicatoren zijn gebaseerd op een niet-representatieve onderzoeksgroep en beperken zich tot parameters uit de klassieke testtheorie. De definitieve versie van de test is IRT[3]-gekalibreerd waarbij de data van een representatieve ijkings-steekproef die gebruikt is voor de normering zijn geanalyseerd. Uit deze analyses die in het volgende hoofdstuk worden gepresenteerd blijkt dat men ook in termen van het gehanteerde IRT-meetmodel de constructie van de test geslaagd mag noemen.

[3] De afkorting IRT staat voor item response theorie; een nadere toelichting volgt in hoofdstuk 4.

4 Het normeringsonderzoek

In dit hoofdstuk komen de volgende onderwerpen aan de orde. Op de eerste plaats bespreken we het steekproefkader. We beschrijven in paragraaf 4.1 de manier waarop we scholen hebben geworven om te komen tot een representatieve steekproef van leerlingen voor verschillende leeftijdsgroepen. Bij de werving van scholen is gelet op de verdeling van scholen naar onderwijstype, regio en mate van verstedelijking.
In een tweede stap is er vervolgens voor gezorgd dat de steekproef ook op leerlingniveau representatief was. Daarbij was het in tweede instantie ook mogelijk om adequate normen per leerjaar te ontwikkelen. Het spreekt vanzelf dat we rekening moesten houden met de verdeling van school- en onderwijstypen naar leeftijd en leerjaar.
Om te komen tot adequate leeftijds- en leerjaarnormen werd gebruik gemaakt van wegingsprocedures op basis van IRT. Daarom beschrijven we in paragraaf 4.2 het gehanteerde meetmodel. Achtereenvolgens beschrijven we in deze paragraaf het marginale multidimensionele OPLM, de wijze waarop de model-parameters werden geschat, de procedures om de modelpassing te realiseren en te beschrijven en ten slotte de wijze van steekproefweging die is toegepast. In paragraaf 4.3 wordt gerapporteerd tot welke resultaten deze procedures hebben geleid in termen van modelpassing en gewogen steekproeven.

4.1 Samenstelling en representativiteit van de normsteekproef

Aanpak van de normering en werving van scholen

Uitgangspunt bij de steekproeftrekking was in eerste instantie te komen tot representatieve leeftijdsgroepen voor de leeftijden 12 tot en met 15 jaar in de eerste drie leerjaren van het VO. De scholen in Nederland vormen de ingang voor de steekproeftrekking. Het steekproefkader is gebaseerd op drie bestanden van CFI[4] waarbij per zogeheten brincode het aantal leerlingen per onderwijsniveau uitgesplitst is naar leeftijd. De bestanden hadden betrekking op het primair onderwijs, voortgezet onderwijs en speciaal onderwijs. Om het steekproefkader te verhelderen met betrekking tot de populatieverdeling naar landelijke regio en urbanisatiegraad zijn deze drie bestanden gekoppeld aan een bestand waarin per brincode de postcode vermeld stond.
Op basis van deze informatie is het steekproefkader berekend voor de populatie 12-, 13-, 14- en 15-jarigen op het moment van de geplande afname, namelijk in de eerste maanden van het kalenderjaar.
Deze berekening is relevant omdat de verdeling van leeftijd over leerjaren in het basisonderwijs en voortgezet onderwijs in de loop van het schooljaar wijzigt. Zo zullen in het begin van een schooljaar in september de meeste leerlingen van 12 jaar in het voortgezet onderwijs zitten (in leerjaar 1) terwijl aan het einde van het schooljaar de meeste 12-jarigen zich in het basisonderwijs bevinden (in groep 8).
De standaard CFI-leeftijd per 31 december (met een teldatum van 1 oktober) moest dus voor de feitelijke afnamedatum worden gecorrigeerd.
Alle Nederlandse scholen werden ingedeeld in 88 verschillende steekproefgroepen, namelijk schooltype (11 niveaus) * regio (4 niveaus) * urbanisatiegraad (2 niveaus). Per school is een toevalsgetal tussen 0 en 1 toegewezen. De scholen met de hoogste toevalsgetallen per steekproefgroep zijn in eerste instantie toegewezen aan de steekproef. Vervolgens is gekeken of steekproefdoel en -resultaat[5] in redelijke mate met elkaar in overeenstemming waren op de belangrijkste achtergrondvariabelen. Door scholen uit de steekproef te halen die op één van de achtergrondvariabelen oververtegenwoordigd waren is een betere matching met de verdeling van de achtergrondvariabelen verkregen. Deze scholen werden vervolgens uitgenodigd deel te nemen aan het normeringsonderzoek met het door Cito vooraf random toegewezen en aangegeven onderwijsniveau, en wel in principe met de volledige leerlingpopulatie van de school (dat wil

[4] CFI staat voor Centrale Financiën Instellingen; tegenwoordig DUO (Dienst Uitvoering Onderwijs). De bestanden werden op ons verzoek door CFI geleverd.

[5] Met steekproefresultaat is hier bedoeld het gerealiseerde steekproefkader: de verzameling van scholen die werd uitgenodigd om aan het normeringsonderzoek deel te nemen.

zeggen met alle leerlingen van 12 tot en met 15 jaar, dus zonder specifieke kinderen uit die leeftijds-categorieën uit te sluiten van deelname).
In de tabellen 4.1 en 4.2 is ten aanzien van het aantal scholen in de steekproef aangegeven in hoeverre steekproefdoel en steekproefresultaat met elkaar in overeenstemming waren op de belangrijke achtergrondvariabelen.

Tabel 4.1 Representativiteit naar onderwijstype (beoogd en gerealiseerd % per leeftijd)

Onderwijstype	12		13		14		15	
	beoogd	werkelijk	beoogd	werkelijk	beoogd	werkelijk	beoogd	werkelijk
Basisonderwijs Stratum 1	28	25	4	4	0	0	0	0
Basisonderwijs Stratum 2	11	15	2	3	0	0	0	0
Basisonderwijs Stratum 3	6	7	2	2	0	0	0	0
VO Brugjaar	25	23	32	27	15	14	3	6
VO pro	0	1	1	5	1	6	1	6
VO vmbo	15	14	35	36	48	44	53	49
VO havo	1	2	7	9	14	15	19	13
VO vwo	6	3	13	6	19	15	21	20
Speciaal basisonderwijs SBaO	4	5	2	2	0	0	0	0
Speciaal onderwijs SO	2	4	1	2	0	1	0	1
VSO	1	1	2	3	2	4	3	6

Tabel 4.2 Representativiteit naar regio en verstedelijking (beoogd en gerealiseerd % per leeftijd)

Regio en verstedelijking	12		13		14		15	
	beoogd	werkelijk	beoogd	werkelijk	beoogd	werkelijk	beoogd	werkelijk
Noord stedelijk	2	4	2	4	2	8	2	6
Noord niet-stedelijk	9	8	8	2	8	3	8	5
Oost stedelijk	8	9	8	3	8	3	9	4
Oost niet-stedelijk	15	13	14	17	14	17	14	17
West stedelijk	31	33	32	32	33	25	33	25
West niet-stedelijk	15	15	13	19	13	18	13	18
Zuid stedelijk	8	7	9	12	9	15	9	13
Zuid niet-stedelijk	14	11	13	12	13	11	13	11

Stratum is een achtergrondkenmerk van een school dat wordt gehanteerd door Cito om ervoor te zorgen dat scholen met verschillende gemiddelde leerlinggewichten in de goede verhouding voorkomen in de steekproef. In de PPON uitgaven van Cito (bijvoorbeeld “Balans van het rekenen-wiskunde onderwijs halverwege de basisschool 5”, Hop, 2012) staat uitgebreid beschreven hoe de stratumindeling tot stand komt. Omdat leerlinggewichten tot stand komen op basis van het opleidingsniveau van de ouders en dit kenmerk een sterke samenhang vertoont met sociaal-economische status kan de variabele stratum op schoolniveau worden opgevat als een indicatieve operationalisatie van de sociaal-economische status. Stratumgegevens zijn alleen voor het basisonderwijs vast te stellen. In totaal werden 287 scholen uitgenodigd om deel te nemen aan het normeringsonderzoek. Helaas was de deelnamebereidheid van scholen in zowel basis- als voortgezet en speciaal onderwijs erg laag. Slechts 50 scholen zegden hun medewerking toe aan het onderzoek. Uiteindelijk waren er 49 scholen waarvan we de data, al dan niet volledig, terug kregen.

Uit analyse van de verzamelde data bleek dat scholen in veel gevallen met andere leerlingen hebben deelgenomen dan was gevraagd. Soms hadden scholen niet met alle gevraagde leerjaren en

onderwijsniveaus met het onderzoek deelgenomen, soms ook met andere dan de gevraagde leerjaren en onderwijsniveaus.
Daarnaast bleken ook relatief weinig leerlingen te hebben deelgenomen van het speciaal onderwijs en met een andere thuistaal dan het Nederlands. Er bleken ook te weinig, met name vertraagde, leerlingen van groep 6 en 7 van het basisonderwijs in de steekproef te zijn opgenomen. Daarom werd voor deze groepen aanvullend onderzoek gedaan. Er werd een extra steekproef getrokken van 150 scholen voor regulier basisonderwijs. Aan deze scholen werd gevraagd deel te nemen met vertraagde leerlingen uit groep 6 en 7 en reguliere leerlingen uit groep 7. Van deze scholen deden er 13 mee met het onderzoek. Daarnaast werden nog eens 33 scholen voor speciaal (basis-)onderwijs uitgenodigd; zes scholen zegden hun medewerking toe, maar van slechts vier kwamen uiteindelijk data retour. Ten slotte bleken, na telefonisch contact, vijf scholen uit Rotterdam en Amsterdam met een hoog percentage allochtone leerlingen bereid mee te doen met het normeringsonderzoek. Deze scholen zijn niet random geworven maar via bestaande contacten van Cito. Van drie van deze scholen kwamen data terug.

De dataverzameling ten behoeve van de normering heeft plaatsgevonden vanaf eind januari 2009 tot en met april 2009. Het aanvullende onderzoek heeft plaatsgevonden tussen februari en mei 2010.
De gemiddelde afnamedatum van het normeringsonderzoek is 15 maart.

Verloop van de dataverzameling

Uitgangspunt bij de dataverzameling was dat de test door de leerlingen nagenoeg geheel zelfstandig gemaakt diende te worden. Tijdens de eerste veldtest in het najaar van 2007 is uitgeprobeerd of de leerlingen de opgaven zonder instructie vooraf konden maken. Uit die observatie en interviews achteraf hebben we kunnen opmaken dat dat het geval was. De opgaven zijn zodanig vormgegeven dat deze min of meer intuïtief juist geinterpreteerd kunnen worden. Daarnaast is veel aandacht besteed aan een heldere en duidelijke instructie die vlak voor de afname door de afnameleider voorgelezen dient te worden. Hiermee is gewaarborgd dat alle leerlingen met dezelfde informatie aan de test zouden beginnen. In die instructie is het centraal invullen van de leerling- en schoolgegevens van het leerlingantwoordblad opgenomen.
De leerlingen kregen tijdens de centrale instructie te horen dat zij de test zelfstandig dienden te maken. Voorafgaand aan elke nieuwe categorie opgaven zijn voorbeelditems opgenomen om duidelijk te maken hoe de opgaven gemaakt moesten worden. Het is belangrijk dat de omstandigheden waaronder de test wordt afgenomen voor alle deelnemende leerlingen zoveel mogelijk hetzelfde te zijn. Dat geldt uiteraard ook voor de afnames in het kader van het normeringsonderzoek. Om een optimale standaardisatie van de afname te realiseren werd in de instructie een aantal richtlijnen opgenomen, die ook in het normerings-onderzoek werden nageleefd.

- De tijd voor een afname werd beperkt tot 50 minuten per afname (50 minuten voor boekje A en 50 minuten voor boekje B). Uit onderzoek vooraf is gebleken dat 93 procent van de leerlingen de test geheel kan afronden binnen die 50 minuten.
- Op de instructiekaart is opgenomen dat de afnameleider halverwege de test dient aan te geven dat de helft van de tijd verstreken is.
- Ook is op de instructiekaart aangegeven welke aanwijzingen een afnameleider mag geven aan de leerlingen. Deze aanwijzingen beperken zich tot het individueel bespreken van de voorbeeldopgaven die voor elke categorie in het afnameboekje staan. Dit is van toepassing op die gevallen waarin een leerling aangeeft dat hij of zij de bedoeling van de opgaven niet begrijpt of wanneer een afnameleider sterke aanwijzingen heeft dat een leerling niet begrijpt wat er van hem wordt verwacht.
- De rol van afnameleider kon worden ingevuld door docenten. Voorwaarde daarbij is dat docenten zich adequaat voorbereiden op de toetsafname. Docenten kregen als richtlijn om de toetsboekjes en de instructiekaart vooraf minimaal één keer zorgvuldig door te nemen.

Verder is het belangrijk te benadrukken dat de opgaven na de eerste proeftoetsing niet meer zijn gewijzigd en ook de presentatievorm van de opgaven, de instructiekaart en de voorbeeldopgaven in de uiteindelijke uitgave exact overeenkomen met die van de normeringsonderzoeken.

Bij het verwerken van de data is gebruik gemaakt van inleesapparatuur. De antwoordbladen die door de apparatuur niet verwerkt konden worden zijn handmatig ingevoerd. Verder zijn de data opgeschoond door alle leerlingen die slechts één boekje hadden gemaakt te verwijderen uit de dataset. Ook leerlingen die per domein meer dan eenderde van de opgaven niet hadden gemaakt (overgeslagen) zijn uit de dataset verwijderd.

Met de materialen die naar de scholen werden verzonden is een voorbeeldbrief meegegaan ter informatie aan de ouders waarin duidelijk is aangegeven dat de ouders bezwaar konden maken tegen deelname aan de intelligentietest. In enkele gevallen (< 15) is door ouders bezwaar gemaakt en hebben die leerlingen de test dan ook niet gemaakt. In de communicatie met de scholen is aangegeven dat de school vooraf duidelijkheid zou moeten verschaffen over de wijze van terugkoppeling van de resultaten aan de ouders. Daarbij is het uitgangspunt altijd geweest dat ouders hoe dan ook uiteindelijk inzage kunnen krijgen in de uitslag van de test. Sommige scholen kozen ervoor om de uitslag van de test mee te geven aan de leerling. Andere scholen kozen ervoor om de uitslag alleen op verzoek kenbaar te maken. De testontwikkelaars bij Cito zijn altijd bereid geweest om toelichting te geven aan de scholen en ouders. Hiervan is door scholen met regelmaat gebruikgemaakt en in het geval van ouders slechts beperkt.

Het moge duidelijk zijn dat de aanvankelijk *op schoolniveau* representatieve steekproef *op leerlingniveau* niet geheel aan onze wensen beantwoordde en dat door middel van een vorm van weging gecorrigeerd zou moeten worden. Die weging bleek overigens ook nodig om adequate, naar *onderwijstype*, *leerjaar* en *leeftijd* van de leerlingen representatieve normgroepen te vormen. We besloten een corrigerende wegings-procedure toe te passen op basis van IRT. Verderop zullen we laten zien hoe deze weging daadwerkelijk werd uitgevoerd. Maar eerst is het nodig om te laten zien welk meetmodel we hebben gehanteerd, hoe we de parameters van het model hebben geschat en wat de resultaten waren in termen van modelpassing.

4.2 Het gehanteerde meetmodel

Waar de afkorting IQ voorheen verwees naar de verhouding tussen mentale en biologische leeftijd is het IQ vandaag de dag een getal dat aangeeft hoe iemand presteert in relatie tot een welbepaalde referentiegroep (i.e., kinderen van een bepaalde leeftijd of in een bepaald leerjaar) en is de score een eenvoudige transformatie van de scoreverdeling in die referentiegroep. Om precies te zijn:

$$IQ(x_+) = \Phi^{-1}[F(X_+ \leq x_+); \mu = 100, \sigma = 15] \qquad (1)$$

Hierbij is X_+ de test-score, $\Phi^{-1}(p\,; \mu = 100, \sigma = 15)$ de inverse van de normale verdelingsfunctie met gemiddelde 100 en standaarddeviatie 15, en $p = F(X_+ \leq x_+)$ het geschatte cumulatieve percentage in de referentiepopulatie. Simpel gesteld is het IQ die waarde onder een normale verdeling waarvan het cumulatieve percentage correspondeert met de gevonden kans in de referentiepopulatie. Het gemiddelde van 100 en de standaarddeviatie van 15 zijn gekozen om historische redenen. Eveneens om historische redenen handhaven we hier de traditionele term IQ (terwijl er van een quotiënt geen sprake is).
De percentages $F(X_+ \leq x_+)$ worden geschat met behulp van de data en de enige voorwaarde voor het correct berekenen van het IQ is dat we een goede schatting hebben van de verdeling van scores in de referentiegroep. Dit laatste blijkt in de praktijk helaas niet eenvoudig te zijn omdat het lastig is om tot een adequate, representatieve steekproef te komen.

We hebben voor de ontwikkeling van de Intelligentietest VO *Item Response Theorie (IRT)* gebruikt om drie redenen:

1. *Validering*: Het IRT-model representeert aannamen waaraan een IQ-test zou moeten voldoen. Als het gepostuleerde model een goede beschrijving geeft van de data draagt dit dus bij aan de validiteit van het instrument.
2. *Betrouwbaarheid*: Met behulp van het IRT-model kunnen we de verdeling van scores bij herhaalde testafname produceren. Hiermee kunnen we de betrouwbaarheid van de test en betrouwbaarheidsintervallen rond het IQ en specifieke deelscores bepalen.

3. *Steekproefweging*: zoals in de vorige paragraaf aangegeven, konden we er niet vanuit gaan dat de scoreverdeling bepaald op basis van de steekproef een goede schatting is van de verdeling in de relevante populatie. We hebben daarom de steekproef moeten herwegen zodanig dat subjecten in verschillende, door achtergrondvariabelen gedefinieerde, groepen voorkomen in de proporties waarin ze in de beoogde populaties voorkomen.

Het IRT-model dat we gebruiken is het *marginale multi-dimensionele One-Parameter Logistic Model (OPLM)*. In dit hoofdstuk beschrijven we hoe we item response theorie hebben gebruikt en wat de resultaten daarvan waren in termen van modelfit en steekproefcorrectie. We doen dit wat uitgebreider dan gebruikelijk omdat we bij de constructie van de intelligentietest gebruik hebben gemaakt van een aantal innovatieve, op IRT gebaseerde technieken. In hoofdstuk 5 rapporteren we daarnaast betrouwbaarheids-gegevens die met behulp van IRT werden berekend en in hoofdstuk 6 een aantal, op basis van IRT afgeleide, validiteitsindicatoren.

4.2.1 Het marginale multidimensionele OPLM

We beschouwen de situatie waarin een steekproef van n personen een vaardigheidstoets of prestatietest[6] (zoals een intelligentietest) heeft gemaakt die bestaat uit k items die goed of fout beantwoord konden worden. De score van persoon p op item i is een binaire toevalsvariabele X_{pi} waarbij $X_{pi} = 1$ als het antwoord goed was en $X_{pi} = 0$ als het antwoord fout was. Een *IRT-model* is een wiskundige beschrijving van de gemeenschappelijke verdeling van de itemantwoorden.

In deze paragraaf beschrijven we het marginale multi-dimensionele OPLM. Daarbij vatten we het OPLM op als een speciaal geval van het Raschmodel. Het Raschmodel, genoemd naar Georg Rasch die het in 1960 introduceerde (o.a. om intelligentietestdata te analyseren), is het simpelste, niet triviale, voorbeeld van een IRT-model en een bouwsteen voor veel complexere modellen, waaronder het OPLM. We bespreken eerst hoe het model is afgeleid uit een aantal redelijke en falsifieerbare aannamen waaraan een prestatietest zou moeten voldoen. Daarna beschrijven we het OPLM als een uitbreiding van het Raschmodel. Vervolgens gaan we in op schatten en het bepalen van de passing van het model.

Monotone, unidimensionale IRT-modellen
De meeste IRT-modellen die voor vaardigheidsmetingen worden gebruikt zijn gebaseerd op de volgende aannamen:

I. *Lokale onafhankelijkheid*: alle responsevariabelen zijn onafhankelijk.
II. *Unidimensionaliteit*: personen met dezelfde vaardigheid hebben dezelfde kans om elk van de items goed te maken.
III. *Monotoniciteit*: hoe groter de vaardigheid, hoe groter de kans op een correct antwoord op elk van de items.

Dat de itemantwoorden onafhankelijk zijn wil zeggen dat elke responsevariabele onafhankelijk Bernoulli verdeeld is. Praktisch gesproken betekent dit dat afkijken of samenwerken niet is toegestaan. De overige aannamen zijn wenselijk geachte eigenschappen voor een vaardigheidsmeting en leggen een vrij sterke structuur op aan de kansen waarmee elke persoon het correcte antwoord op elk van de items vindt. Unidimensionaliteit impliceert dat personen gekarakteriseerd kunnen worden door één getal dat hun vaardigheid representeert. Deze vaardigheid is te interpreteren als de score op een oneindig lange test. Uit monotoniciteit volgt vervolgens dat elk van de kansen een monotoon stijgende functie is van de vaardigheid; i.e., voor alle items i geldt dat $P(X_{pi} = 1) = \phi_i(\theta_p)$, waarbij θ_p de vaardigheid is van persoon p. De functie ϕ_i is onafhankelijk van de persoon en wordt de *itemkarakteristieke curve* genoemd.

[6] Deze passage is in principe van toepassing op elk instrument, toets of test, waarbij het onderzochte individu geacht wordt te laten zien waartoe hij in staat is. In het vervolg gebruiken we hier het woord vaardigheidsmeting in deze algemene betekenis.

Sufficiëntie van de testscore
De testscore (typisch het aantal goede antwoorden) is een statistiek op basis waarvan beslissingen worden genomen over personen. Het is van groot belang dat deze score alle informatie bevat over de vaardigheid die in het geding is. We maken daarom de volgende aanname:

IV. *Sufficiëntie van de testscore*: De testscore bevat alle informatie over de vaardigheid van een persoon.

Sufficiëntie wil zeggen dat we niets meer bijleren over de vaardigheid als we de testscore eenmaal hebben geobserveerd. Formeel betekent dit dat, conditioneel op de geobserveerde score, de verdeling van de data onafhankelijk is van vaardigheid.

Het Raschmodel
Als we aannemen dat het aantal goede antwoorden op de test sufficiënt is dan volgt het Raschmodel.

V. *(Rasch) Het aantal goede antwoorden is een sufficiënte statistiek voor vaardigheid*

Dit is een unieke eigenschap van het Raschmodel en rechtvaardigt het gebruik van het aantal goede antwoorden als een testscore. Als deze aanname houdt, volgt dat

$$X_{pi}|\theta_p, \delta_i \sim Bernoulli(\Psi(\theta_p - \delta_i)) \quad (2)$$

waarin θ_p de vaardigheid representeert van persoon p, δ_i de moeilijkheid van het item en Ψ de logistische functie $\Psi(x) = \exp(x)/(1 + \exp(x))$. We zien dat de kans op een goed antwoord uitsluitend afhangt van het verschil tussen de vaardigheid van een persoon en de moeilijkheid van het item. Hoe groter de vaardigheid van een persoon relatief ten opzichte van de moeilijkheid van het item, hoe groter de kans op een correct antwoord. Merk op dat $\theta_p - \delta_i = (\theta_p - c) - (\delta_i - c)$ voor een willekeurige constante c. Dat wil zeggen dat alleen de relatieve vaardigheid/moeilijkheid uniek bepaald is. Het nulpunt van de vaardigheidsschaal is onbepaald en wordt willekeurig gefixeerd op bijvoorbeeld één van de itemmoeilijkheden.

Het OPLM
Onder het Raschmodel is het aantal goede antwoorden $X_+ = \sum_i X_i$ een sufficiënte statistiek voor vaardigheid. Dat betekent dat het er niet toe doet *welke* items een persoon goed gemaakt heeft. Alleen het aantal telt. In de praktijk zijn er echter situaties waarbij het aantal punten dat wordt verdiend met een goed antwoord verschilt over items. Dat is bijvoorbeeld zo wanneer a_i punten worden gegeven voor een goed antwoord op item i en de a_i niet voor alle items gelijk zijn. De toetsscore is dan $X_+ = \sum_i a_i X_i$. Als deze testscore sufficiënt is dan volgt het OPLM.

VI. *(OPLM) Het totale aantal punten is een sufficiënte statistiek voor vaardigheid*

Onder het OPLM geldt dat:

$$X_{pi}|\theta_p, \delta_i, a_i \sim Bernoulli(\Psi(a_i[\theta_p - \delta_i])) \quad (3)$$

De constanten a_i worden ook wel discriminatie-indices genoemd vanwege het feit dat de kans om het item goed te beantwoorden sneller verandert als functie van vaardigheid naarmate de discriminatie-index een hogere waarde heeft. Het Raschmodel is een speciaal geval van het OPLM waarbij $a_i = 1$ voor alle items i.

Het OPLM is door Verhelst en Eggen (1989) voorgesteld als een uitbreiding van het Raschmodel. Een gepubliceerde technische uiteenzetting van het model is te vinden in Verhelst en Glas (1995). Het bijbehorende computerprogramma wordt beschreven in Verhelst, Glas en Verstralen (1995).

Marginale IRT-modellen
In de huidige toepassing nemen we aan dat de subjecten (personen, kinderen, leerlingen) een steekproef vormen uit een wel-gedefinieerde populatie. Dit is de vijfde aanname:

VII. *Random personen*: Vaardigheid is een toevalsvariabele.

De verdeling van vaardigheid is afhankelijk van de groep waartoe een persoon behoort; bijvoorbeeld 12-jarigen. De groep wordt gedefinieerd door de waarde op een vector discrete achtergrond variabele $\boldsymbol{t}$.

Alle aannamen samen induceren een statistisch model voor de verdeling van de data conditioneel op de achtergrond variabele $\boldsymbol{t}$:

$$P(\boldsymbol{X} = x) = \prod_p P\left(\boldsymbol{X}_p = \boldsymbol{x}_p | \boldsymbol{t}_p\right) \tag{4}$$

$$= \prod_p \int_{-\infty}^{\infty} \prod_i P\left(X_{pi} = x_{pi} | \theta, \delta_i\right) f(\theta | \boldsymbol{t}_p) d\theta \tag{5}$$

Dit is een unidimensioneel, marginaal IRT-model: een marginaal Raschmodel als we het Raschmodel gebruiken en een marginaal OPLM als we OPLM gebruiken. Merk op dat we impliciet aannemen dat de op vaardigheid conditionele verdeling onafhankelijk is van de waarde van de achtergrondvariabele.
Deze aanname heet meetinvariantie.

VIII. *Meetinvariantie*: Het IRT-model is onafhankelijk van de waarde van de achtergrondvariabele.

Meetinvariantie betekent dat hetzelfde IRT-model houdt in elke groep. Informeel kunnen we zeggen dat de test dezelfde vaardigheid meet in bijvoorbeeld de groep 12-jarigen en de groep 14-jarigen. Als deze aanname geschonden wordt, is er sprake van “Differential Item Functioning” of kortweg DIF. Het kan overigens wel zo zijn dat de 14-jarigen vaardiger zijn of juist minder vaardig dan de 12-jarigen terwijl er toch geen sprake is van DIF.

Multidimensionaliteit
Een IQ-test ter vaststelling van de algemene intelligentie bestaat per definitie uit een verzameling heterogene opgaven (zie hoofdstuk 1). In de Intelligentietest VO onderscheiden we een aantal deeltaken waarvan we op inhoudelijke gronden aannemen dat ze elk een verschillende vaardigheid meten.
We veronderstellen een marginaal OPLM voor elke deeltaak, respectievelijk deelvaardigheid. Voor de vaardigheden veronderstellen we dat ze een gemeenschappelijke verdeling hebben. Voor twee deeltaken wordt het model dan

$$P(\boldsymbol{x}, \boldsymbol{z} | \boldsymbol{t}) = \prod_p \int_{\infty}^{\infty} \int_{\infty}^{\infty} P\left(\boldsymbol{x}_p | \theta^{(1)}\right) P(\boldsymbol{z}_p | \theta^{(2)}) f(\theta^{(1)}, \theta^{(2)} | \boldsymbol{t}_p) d\theta^{(1)} d\theta^{(2)} \tag{6}$$

waarin $\boldsymbol{x}$ en $\boldsymbol{z}$ itemantwoorden representeren op verschillende deeltaken, en $\theta^{(1)}$ en $\theta^{(2)}$ de corresponderende vaardigheden. In al onze analyses veronderstellen we dat de vaardigheden multivariaat normaal verdeeld zijn.

4.2.2 Het schatten van de modelparameters

In deze paragraaf beschrijven we kort hoe we met behulp van de data de parameters van het model schatten. De itemparameters kunnen per deeltaak geschat worden met *conditional maximum likelihood*. Vervolgens baseren we inferenties over vaardigheid op trekkingen uit de a posteriori verdeling van

vaardigheid. Deze “verdeel-en-heers strategie“ waarbij eerst de items worden geanalyseerd en daarna de verdeling van vaardigheid wordt mogelijk gemaakt door de aanwezigheid van een sufficiënte statistiek.

Het schatten van item-moeilijkheid
De verdeling van de data gegeven de geobserveerde waarden voor de sufficiënte statistiek (i.e., de gewogen of ongewogen somscores) hangt niet af van de vaardigheid. We kunnen het model daarom schijven als:

$$\prod_p P(\boldsymbol{x}_p|\boldsymbol{t}_p) = \prod_p \int_{-\infty}^{\infty} P(\boldsymbol{x}_p|x_{p+}) P(x_{p+}|\theta) f(\theta|\boldsymbol{t}_p) d\theta \quad (7)$$

$$= \left[\prod_p P(\boldsymbol{x}_p|x_{p+})\right] \prod_p \int_{-\infty}^{\infty} P(x_{p+}|\theta) f(\theta|\boldsymbol{t}_p) d\theta \quad (8)$$

De eerste factor is de verdeling conditioneel op de geobserveerde scores. De tweede term is $\prod_p P(x_{p+}) = P(\boldsymbol{x}_{.+})$; de marginale verdeling van de scores: $\boldsymbol{x}_{.+} = (x_{1+}, x_{2+}, \ldots, x_{n+})$. Als we de tweede term gelijk stellen aan de geobserveerde proporties in de steekproef dan krijgen we een *extended Rasch/OPLM*.

Als we het *extended Raschmodel* aannemen, kunnen we voor het schatten van de itemparameters de scoreverdeling negeren en ons baseren op de conditionele verdeling. Deze verdeling hangt niet af van vaardigheid, alleen van de item-moeilijkheidsparameters.

Voor het schatten maken we gebruik van de methode van *maximum likelihood (ML)*. Dit is een standaard schattingsmethode waarbij we de conditionele verdeling beschouwen als een functie van de parameters (de *likelihood*-functie). De ML-schatters zijn die waarden van de parameters die de *likelihood*-functie maximaliseren, of met andere woorden de waarden waarvoor de kans op de geobserveerde data zo groot mogelijk is. Wanneer we de conditionele verdeling gebruiken spreken we van *conditional maximum likelihood (CML).* Een belangrijk voordeel van CML is dat itemparameters consistent worden geschat ook wanneer de verdeling van vaardigheid verkeerd is gespecificeerd. CML is de standaard schattingsmethode. De technische details van het schatten staan uitgelegd in onder andere Bechger en Maris (2010).

Discriminatie-indices
OPLM veronderstelt dat we de discriminatie-indices kennen. In de praktijk is dit echter vaak niet het geval en moeten we de discriminatie-indices noodgedwongen schatten. Er zijn twee manieren waarop dat kan worden gerealiseerd:

1. We schatten een Raschmodel en passen de discriminatie-index van een item aan wanneer het Raschmodel niet past (zie hieronder).
2. We beschouwen de discriminatie-indices als onbekende parameters en schatten ze.

Een model waarin de discriminatie-indices onbekende parameters zijn wordt het *twee-parameter model (2PL)* genoemd. In dit model is er geen sufficiënte statistiek voor de vaardigheid en kunnen we dus geen gebruikmaken van CML. We moeten voor het schatten dus gebruik maken van *marginal maximum likelihood (MML)* waarbij de *likelihood* gebaseerd is op de marginale verdeling van de data (6).

Het schatten van de discriminatieparameters onder het 2PL is complex en op dit moment beschikken we over software voor het geval met een enkele normaal verdeelde vaardigheid (zie Maris en Bechger, 2010). Software voor meer correcte schatting is experimenteel en daarom niet toegepast in dit project.

Inferenties met betrekking tot vaardigheid
Inferenties over vaardigheid baseren we op de verdeling van vaardigheid conditioneel op de geobserveerde ongewogen score. Met twee deeltaken is deze verdeling

$$f(\theta^{(1)}, \theta^{(2)}|x_{p+}, z_{p+}, \boldsymbol{t}_p) \propto P(x_{p+}|\theta) P(z_{p+}|\theta) f(\theta^{(1)}, \theta^{(2)}|\boldsymbol{t}_p) \quad (9)$$

Deze verdeling; *de posterior verdeling van vaardigheid*, representeert de onzekerheid ten aanzien van de vaardigheid van een persoon met x_{p+} correcte itemantwoorden.

We gebruiken de ongewogen scores hier omdat dit beter aansluit bij de praktijk. Dit is volledig efficiënt wanneer de items voldoen aan het Raschmodel. Onder het OPLM verliezen we informatie wanneer we de ongewogen somscore gebruiken naarmate de items meer verschillen in discriminatie. Dit informatieverlies is in de praktijk beperkt. Zowel onder het Raschmodel als onder het OPLM geldt dat de posteriori verdeling geordend is met de score. Alleen onder het Raschmodel geldt dat de posterior verdeling van twee personen met hetzelfde aantal goede antwoorden precies gelijk is. Dit is opnieuw een reden om inferenties te baseren op een sufficiënte statistiek.

De posteriori verdeling van vaardigheid is niet analytisch te bepalen. Om die reden gebruiken we computer simulatie en genereren we onafhankelijke identiek verdeelde trekkingen uit deze verdeling met behulp van het *Conditional-Composition (CC) algoritme* beschreven in Marsman, Maris, Bechger en Glas (2011). Deze trekkingen worden in de literatuur aangeduid als *plausible values (PV)*. PVs worden o.a. gebruikt in projecten als NAEP en PISA (zie bijvoorbeeld Mislevy, Johnson en Muraki, 1992). Het gebruik van PVs is een routine aangelegenheid. Het CC-algoritme is recent ontwikkeld maar naar onze mening voldoende onderzocht om te worden toegepast. Het algoritme is ook toegepast in *The first European survey of language competences*. Merk op dat we bij het genereren van PVs de itemparameters vastzetten op hun CML-schattingen.

PVs worden gebruikt voor twee doelen:
1. Het schatten van de posteriori verdeling van vaardigheid van een persoon. Het gemiddelde van de verdeling van PVs van een persoon is de zogenoemde *expected a posteriori (EAP)* schatter.
2. Het schatten van vaardigheidsverdelingen in de populatie.

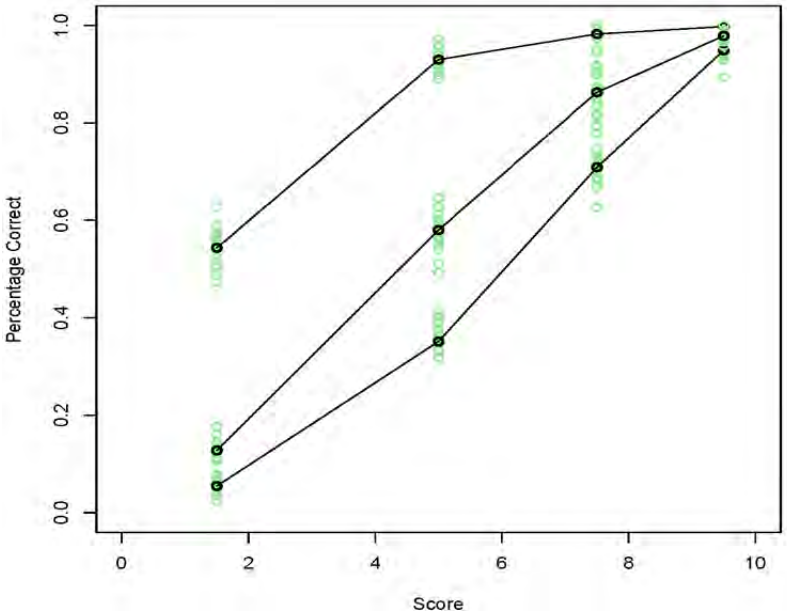
In het kader van de constructie van deze intelligentietest was met name het tweede doel belangrijk. PVs hebben een opmerkelijke eigenschap waardoor ze voor dit doel heel nuttig zijn. Er kan namelijk worden bewezen dat de marginale verdeling van de PVs (dat wil zeggen de verdeling van PVs over personen) een consistente (en niet-parametrische) schatter is van de verdeling van vaardigheid als het aantal personen in de steekproef toeneemt (Marsman, Maris, Bechger en Glas, *In voorbereiding*). Opmerkelijk is dat dit ook geldt wanneer de vaardigheidsverdeling niet correct gespecificeerd is, dus ook wanneer de vaardigheids-verdeling niet multivariaat normaal is. Kortweg komt het erop neer dat de gepostuleerde vaardigheids-verdeling in het model fungeert als a priori verdeling. Naarmate we over meer data beschikken zal het IRT-model ervoor zorgen dat de posteriori verdeling van vaardigheid steeds meer gaat lijken op de correcte posteriori verdeling: Een resultaat dat bekend staat als het Bernstein-von Mises theorema. Voorwaarde is wel dat het IRT-model correct is.

4.2.3 Modelpassing

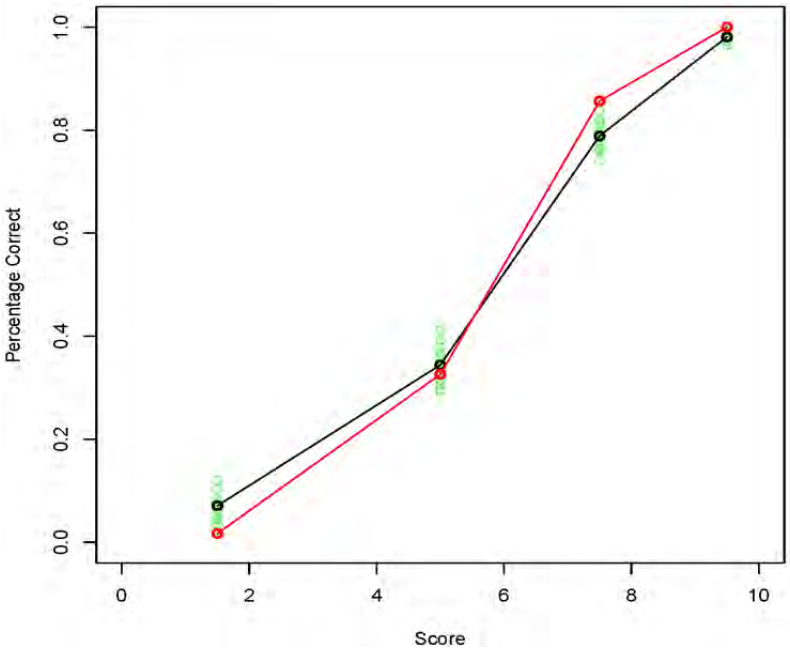
Alle uitspraken die we doen met behulp van ons IRT-model gelden in principe alleen als het model houdt. Het is dus van groot belang te kunnen vaststellen of het model houdt of niet.
Modelpassing wordt onderzocht door te kijken naar de overeenkomst tussen resultaten op basis van het model en uit geobserveerde (dat wil zeggen via testafname verkregen) data. Naast informele methoden, zoals visuele inspectie van grafische weergaven, wordt door Cito gewoonlijk gebruik gemaakt van een statistische toets (i.e. de zogeheten S-toets) die werd ontwikkeld door Verhelst en Glas (1995).
Deze toetsstatistiek is een functie van verschillen in geobserveerde en door het model voorspelde aantallen correcte antwoorden in groepen respondenten met ongeveer dezelfde score.
Ter illustratie toont figuur 4.1 voor drie items de zogeheten *empirische itemresponse-curve*. Iedere curve is gemaakt door respondenten te verdelen in vier groepjes met ongeveer dezelfde score. Vervolgens is voor elk groepje bepaald welk percentage het correcte antwoord gaf. Deze percentages zijn vervolgens afgezet tegen de gemiddelde score in elk groepje. De lijn geeft aan wat we onder het Raschmodel verwachten.
De (groene) bolletjes rond ieder percentage geeft aan wat de variatie is die we mogen verwachten over

steekproeven (in dit geval van $n = 600$). Als de geobserveerde percentages meer afwijken dan verwacht mag worden ten gevolge van steekproefvariantie dan verwerpen we de hypothese dat het model past. Figuur 4.2 geeft een voorbeeld van een OPLM item. We zien aan het patroon van afwijkingen dat dit item meer discrimineert dan onder het Raschmodel wordt verondersteld. Door de afwijkingen slim te combineren over score-groepen en/of items verkrijgen we een statistiek waarmee de *fit* van het model kan worden getoetst: de R1c-toets. Voor technische details verwijzen we de lezer naar Verhelst en Glas (1995).

Figuur 4.1 Voorbeelden van empirische itemresponse-curves voor drie items (toelichting in de tekst)



Figuur 4.2 Empirische itemresponse curves voor een item dat anders discrimineert dan voorspeld onder het Raschmodel



4.2.4 Betrouwbaarheidsintervallen

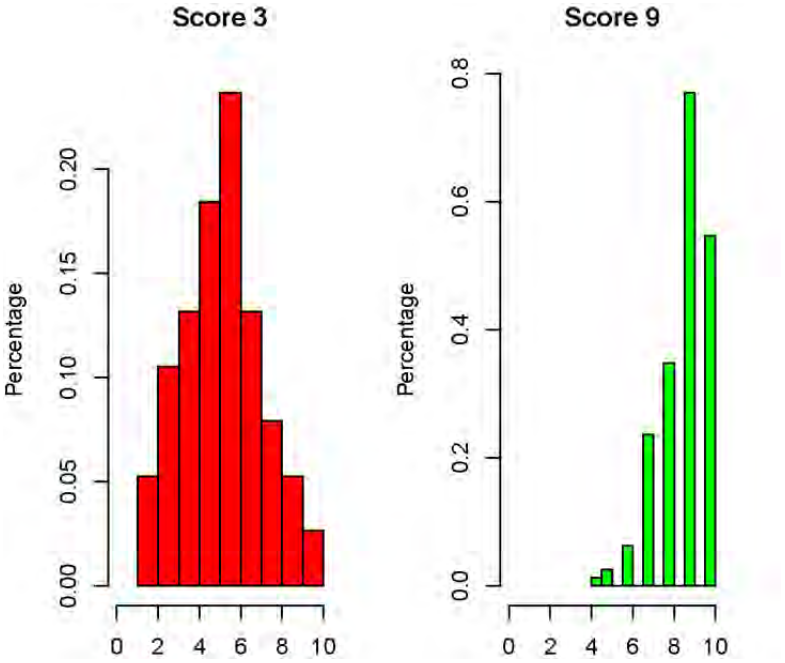
In de inleiding op deze paragraaf (4.2) hebben we aangegeven dat we IRT-procedures niet alleen hebben toegepast om redenen van steekproefweging, maar ook omwille van een adequate schatting van de betrouwbaarheid. Voordat we in paragraaf 4.2.5 onze argumentatie met betrekking tot de steekproefweging vervolgen, gaan we nu eerst kort in op de procedures die we hebben gehanteerd om de betrouwbaarheid te bepalen. Het resultaat van de analyses bespreken we in hoofdstuk 5 waarin betrouwbaarheid en nauwkeurigheid van de Intelligentietest VO centraal staan.

Op de leerlingrapporten worden gegevens over de algemene intelligentie van de leerling vermeld in de vorm van een leeftijd-IQ en een leerjaar-IQ. Daarnaast vermeldt het leerlingrapport het 80%-betrouwbaarheidsinterval voor deze IQ's. In feite gaat onze voorkeur uit naar het rapporteren van een intervalschatting van het IQ. Wanneer we de IQ's (alleen) als een getal zouden rapporteren, gaan we immers voorbij aan de onzekerheid ten gevolge van a) het feit dat de test niet perfect betrouwbaar is en b) het feit dat de scoreverdeling in de referentiepopulatie geschat is. De intervalschattingen zijn bedoeld om deze onzekerheid correct te rapporteren. Aansluitend bij de bestaande praktijk rapporteren we zowel het IQ als het betrouwbaarheidsinterval. Hier gaan we vooral in op de bepaling van de betrouwbaarheids-intervallen.

De betrouwbaarheidsintervallen zijn gebaseerd op de a posteriori voorspelde ("posterior predictive") verdeling van de testscores. Dit is de verdeling van scores bij een tweede (hypothetische) afname van de test geschat op basis van de in het normeringsonderzoek verzamelde data. Hoe we deze verdeling gebruiken om betrouwbaarheidsintervallen te bepalen, zullen we illustreren aan de hand van een voorbeeld.

Voorbeeld
Stel dat de test willekeurig wordt verdeeld in twee helften. We nemen de eerste helft af en observeren een score x_{p+}. Nu nemen we ook de tweede helft af en we zien dat leerlingen die op de eerste helft dezelfde score haalden, nu verschillend scoren. Het bereik van de scores op de tweede test is een redelijke maat voor de onzekerheid ten aanzien van de score op de tweede helft, gegeven het feit dat we weten dat de betreffende leerlingen een score x_{p+} hadden op de eerste helft. Ter illustratie nemen we de schaal FC die bestaat uit 20 items; we gebruiken de gegevens van de 12-jarigen. De histogrammen in figuur 4.3 laten zien dat de scores op de tweede helft variëren, ook onder mensen die op de eerste helft dezelfde score hadden. We kunnen op basis van de eerste helft dus niet precies voorspellen wat mensen op de tweede helft gaan scoren. Dit is het gevolg van meetfout. Daarnaast valt op dat a) de scores op de tweede helft tenderen wat hoger te zijn als de score op de eerste helft hoger is, en b) dat het rechter histogram scheef is naar links. Dit laatste treedt op doordat de maximale score op de tweede testhelft 10 was.

Figuur 4.3 Histogram van scores op de tweede testhelft gegeven een score van 3 (links) of een score van 9 (rechts) op de eerste testhelft (zie beschrijving voorbeeld in de tekst).



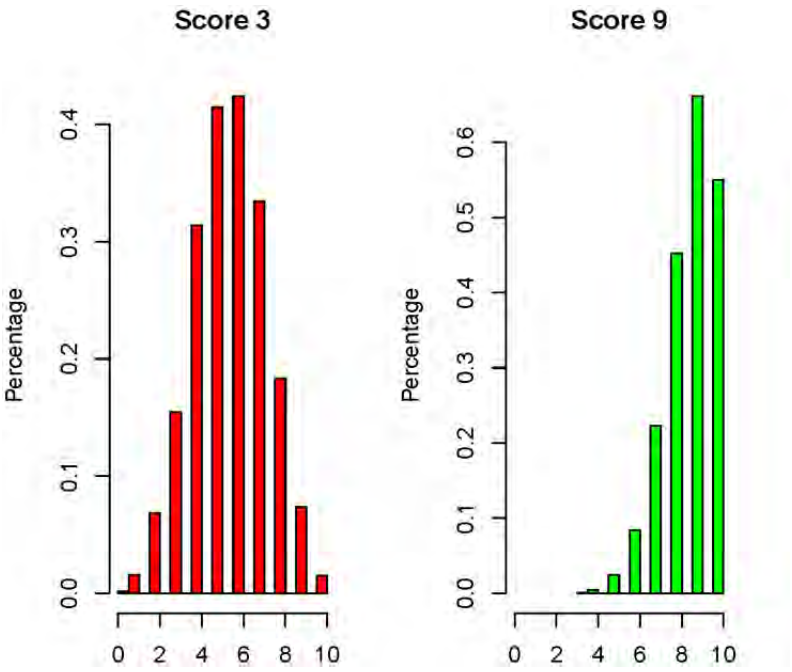
We observeren een (ongewogen) score x_{p+} en daarmee genereren we scores X_+^* uit de verdeling $P(X_+^*|X_+ = x_{p+})$. Onder de assumptie dat de scores onafhankelijk zijn gegeven de vaardigheid kan deze verdeling als volgt worden geschreven:

$$P(X_+^*|X_+ = x_{p+}) = \int_{-\infty}^{\infty} P\,(X_+^*|\theta) f(\theta|X_+ = x_{p+}) d\theta \qquad (10)$$

Hierbij is $P(X_+^*|\theta)$ de verdeling van de scores onder het IRT-model. Om scores te genereren gebruiken we de *compositiemethode* (Tanner, 1993, 3.3.2): trek op basis van toeval een PV en genereer daarmee een testscore. We doen dit zo vaak als we willen. De aldus getrokken testscores zijn een *i.i.d.*[7] trekking uit de verdeling $P(X_+^*|X_+ = x_{p+})$.

[7] 'i.i.d' staat voor 'independent (and) identically distributed'

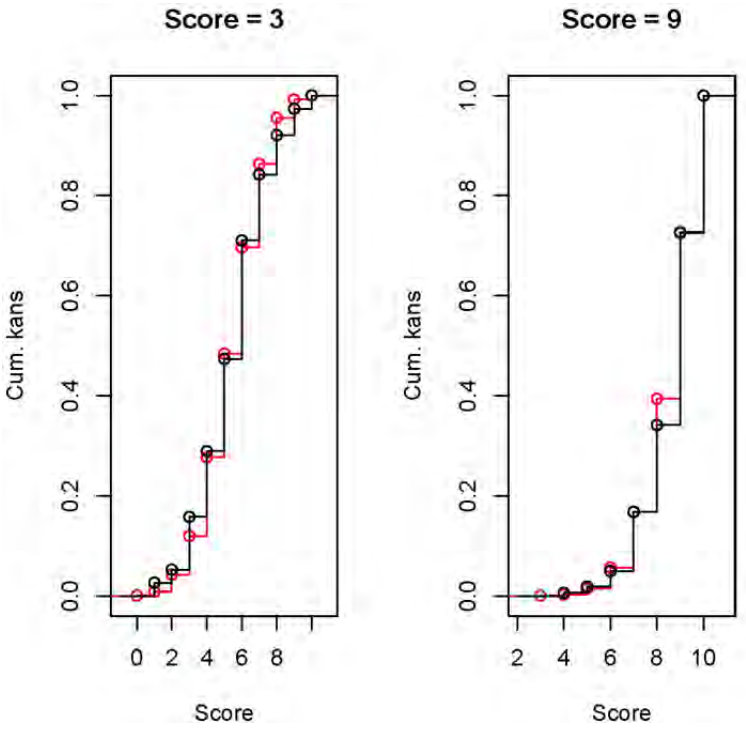
Figuur 4.4 *Histogram van scores op de tweede testhelft gegeven een score van 3 (links) of een score van 9 (rechts) op de eerste testhelft. Ditmaal geschat met behulp van het Rasch-model met 10.000 gerepliceerde scores (zie beschrijving voorbeeld in de tekst).*



Vervolg van het voorbeeld. *Ter illustratie geeft figuur 4.4 de histogrammen voor scores op de tweede testhelft zoals geschat met behulp van de procedure die we in deze paragraaf beschrijven. Indien het Rasch-model past verwachten we dat de scores op de tweede testhelft die we onder het Rasch-model produceren overeen komen met de scores die we hebben geobserveerd. Figuur 4.5 laat zien dat dit in dit geval met hoge waarschijnlijkheid zo is en biedt evidentie dat het Rasch-model op de data past.*
Op basis van figuur 4.5 hebben we geen reden om aan te nemen dat het Rasch-model niet houdt.
De predicties van het model komen immers op het oog overeen met de observaties. Anders gezegd: we maken geen fouten door ons te baseren op de door het model geproduceerde scoreverdeling.

Om een a%-betrouwbaarheidsinterval te bepalen, bepalen we tussen welke grenzen de gerepliceerde scores met a% waarschijnlijkheid liggen. Als we de testscores transformeren naar IQ's verkrijgen we daarmee een betrouwbaarheidsinterval voor het IQ. Het voordeel van deze Bayesiaanse betrouwbaarheidsintervallen boven de meer gebruikelijke, frequentistische betrouwbaarheidsintervallen is dat ze gemakkelijker te interpreteren zijn. Verder is het, doordat we scores genereren, gemakkelijk om betrouwbaarheidsintervallen van scores die van deze testscores zijn afgeleid (zoals het IQ) te bepalen.

Figuur 4.5 Geobserveerde (zwart) en berekende (rood) verdeling van scores op de tweede testhelft (zie beschrijving voorbeeld in de tekst).



Merk op dat de betrouwbaarheidsintervallen de onzekerheid ten gevolge van steekproefvariatie en de onbetrouwbaarheid van de toets correct verdisconteren. Dit geldt niet voor de onzekerheid ten aanzien van het IRT-model. De item-parameters worden immers bekend verondersteld. De onzekerheid is echter niet heel groot en in de praktijk verwaarloosbaar. In alle gevallen is de standaardfout van de moeilijkheden kleiner dan 0,2 keer de standaarddeviatie van de vaardigheid. Bij toekomstige uitgaven zullen we een volledig Bayesiaanse schattingsprocedure uitwerken en gebruiken waarmee deze kleine onzekerheid wel correct wordt meegenomen.

Met behulp van de hierboven beschreven schattingsmethode kan voor alle IQ-waarden afzonderlijk het bijbehorende betrouwbaarheidsinterval worden berekend. In hoofdstuk 5 zullen we de resultaten met betrekking tot de bepaling van de betrouwbaarheid en de berekening van betrouwbaarheidsintervallen rapporteren.

4.2.5 Steekproefweging

Het idee achter weging

Voor het vaststellen van adequate normen is een “goede” steekproef onontbeerlijk. Een goede steekproef is een steekproef die op zo’n manier wordt getrokken dat we met de gegevens van de steekproef een consistente schatter kunnen construeren van de scoreverdeling in elke referentiegroep waarvoor we normen nodig hebben (in ons geval leeftijd- en leerjaargroepen). De huidige steekproef werd weliswaar geworven op basis van de verdeling van Nederlandse scholen op een aantal belangrijke achtergrond-variabelen (regio en mate van verstedelijking), maar dit volstaat niet om representativiteit op leerlingniveau te kunnen garanderen. Bovendien moesten hiaten in de dataverzameling in tweede instantie via aanvullend onderzoek worden gedicht. In deze situatie hebben we onze toevlucht moeten nemen tot weging achteraf op de belangrijkste variabelen leeftijd, leerjaar en schooltype (onderwijsniveau). Daarbij konden we gebruikmaken van het feit dat de test IRT-gekalibreerd was.

Het idee is in principe vrij eenvoudig. Stel, we hebben goede steekproeven van 14-jarigen in elk schooltype maar we zijn er niet zeker van dat de schooltypen in de juiste proporties in de steekproef zitten. Als er, bijvoorbeeld, te veel vwo- leerlingen in de steekproef zitten, dan bestaat de kans dat de gemiddelde score te hoog is. Als gevolg daarvan krijgt een 14-jarige havo-leerling bij toepassing van een normtabel die is gebaseerd op deze steekproef een te laag IQ. Hij of zij wordt immers vergeleken met een referentie-populatie die vooral uit vwo-leerlingen bestaat en dat is niet de referentiepopulatie die we beogen.
Nu, wanneer we kunnen achterhalen hoeveel 14-jarige leerlingen er in de populatie in elk schooltype aanwezig zijn, kunnen we daarmee de proporties in de steekproef “aanpassen”. Dat kan door elke observatie een gewicht te geven bij het berekenen van de verdeling, of door een steekproef te trekken uit de data zodanig dat de percentages leerlingen in elk schooltype in de nieuwe steekproef overeenkomen met de percentages in de populatie. De eerste methode is gebruikelijk en wordt bijvoorbeeld toegepast binnen projecten als PISA. We hebben hier om praktische redenen gekozen voor de tweede methode. Voor het gemak blijven we in het vervolg spreken van steekproefweging.

Resumerend: we veronderstellen dat we een goede steekproef hebben van een groep personen die is gedefinieerd door hun waarden op een vector met (relevante) achtergrondvariabelen zoals bijvoorbeeld schooltype, leeftijd en leerjaar. Als we weten welk percentage van de referentiepopulatie tot elke groep behoort dan kunnen we de steekproefproporties gelijk maken aan de populatieproporties.

Weging achteraf werkt alleen wanneer aan twee voorwaarden is voldaan:
1. Ten eerste doen we de aanname dat we een goede steekproef hebben van elke groep, dat wil zeggen. voor alle (combinaties van) achtergrondvariabelen.
2. Ten tweede nemen we aan dat de achtergrondvariabelen die worden gebruikt samen een ‘voldoende’ set vormen. Dat wil zeggen, dat *alle* variabelen die een rol hebben gespeeld bij het al dan niet deelnemen van scholen en leerlingen aan het onderzoek zijn opgenomen.

Of deze aannamen houden kunnen we hier niet met zekerheid zeggen. Met name de tweede assumptie is in de praktijk niet te toetsen. Ten slotte kunnen we, praktisch gezien, alleen wegen wanneer we beschikken over de populatiegegevens, wat niet altijd het geval is. Dit punt komt aan de orde wanneer we de resultaten bespreken.

Wegen van vaardigheidsverdelingen
De PVs vormen een steekproef uit de vaardigheidsverdeling. Deze verdeling is echter niet de juiste wanneer we geen goede steekproef hebben van personen. Om de steekproef representatief te maken trekken we een steekproef uit de steekproef zodanig dat het aantal personen in elke groep in de steekproef overeenkomt met de populatieproporties.

De vector met achtergrondvariabelen geven we aan met T. De vaardigheidsverdeling is te schrijven als:

$$f(\theta) = \sum_{t} f\,(\theta|T = t)P(T = t)$$
$$= \sum_{t}\left(\sum_{x_+} f\,(\theta|x_+, t)P(x_+|T = t)\right)P(T = t)$$

De expressie beschrijft bijna letterlijk wat we moeten doen. Eerst trekken we een combinatie van de achtergrondvariabelen met de populatieproporties als kansen. Daarna kiezen we willekeurig een somscore in die groep en genereren een PV bij deze somscore. Het komt erop neer dat we, met teruglegging, een steekproef uit de data trekken zodanig dat alle waarden van de achtergrondvariabele in de juiste proportie voorkomen. De verdeling van de PVs van deze steekproef is een gewogen schatter van de populatieverdeling.

Voorspelde scores en steekproefweging
De verdeling van voorspelde scores X_+^* , gegeven dat de geobserveerde score gelijk is aan $X_+ = x_+$, is te schrijven als:

$$P(X_+^*|X_+ = x_+) = \frac{P(X_+^*, X_+ = x_+)}{P(X_+ = x_+)} \tag{11}$$

Stel, we hebben te maken met schooltype als achtergrondvariabele T. De verdeling $P(X_+^*, X_+ = x_+)$ is te schrijven als:

$$P(X_+^*, X_+ = x_+) = \sum_t P\,(X_+^*, X_+ = x_+|T = t)P(T = t) \tag{12}$$

Zoals eerder uiteengezet houdt steekproefweging het volgende in. We veronderstellen dat we een representatieve steekproef hebben van personen in elk schooltype en nemen aan dat we de schooltypen niet in de juiste verhouding hebben geobserveerd in de steekproef. We gebruiken voor $P(T = t)$ dan niet de geobserveerde proporties maar de proporties zoals ze gelden in de populatie.

De verdeling van de gerepliceerde scores schatten we door scores met behulp van de computer te genereren. Daarbij houden we rekening met de achtergrondvariabele T. Om te zien hoe dit in zijn werk gaat is het goed de verdeling van $P(X_+^*, X_+ = x_+)$ nader te beschouwen:

$$\begin{aligned} P(X_+^*, X_+ = x_+|T = t) &= \int P\,(X_+^*, X_+ = x_+|\theta) f(\theta|T = t) d\theta \\ &\qquad \Downarrow X^* \text{ en } X_+ \text{ zijn onafh. gegeven } \theta \\ &= \int P\,(X_+^*|\theta) P(X_+ = x_+|\theta) f(\theta|T = t) d\theta \\ &= \int P\,(X_+^*|\theta) f(\theta|X_+ = x_+, T = t) d\theta P(X_+ = x_+|T = t) \end{aligned}$$

De laatste vergelijking volgt uit de voorlaatste omdat

$$P(X_+ = x_+|\theta) f(\theta|T = t) = P(X_+ = x_+|T = t) f(\theta|X_+ = x_+, T = t)$$

onder aanname van meetinvariantie. Substitueren we de expressie voor $P(X_+^*, X_+ = x_+|T = t)$ in (12) dan zien we hoe we - met behulp van de compositie methode - uit de verdeling van voorspelde scores moeten trekken. Kort samengevat komt de procedure erop neer dat we de gewogen steekproef van PVs gebruiken om scores te genereren.

4.3 Modelpassing en steekproefweging: resultaten

Nadat we in de vorige paragraaf de theorie achter de toegepaste procedures hebben besproken gaan we hier in op de resultaten. We bespreken eerst steekproefweging. De normering is vervolgens gebaseerd op de gewogen steekproef zoals die is weergegeven in tabel 4.7.
Gegeven de beschikbare data is er uiteindelijk voor gekozen om leeftijdnomen te ontwikkelen voor 11-, 12-, 13- en 14-jarigen in het voortgezet onderwijs. De data volstonden niet om normen te genereren voor de groep 15-jarigen omdat we over te weinig leerlingen van 15 jaar uit leerjaar 4 van het voortgezet onderwijs beschikten. Wel bleken we, op grond van aanvullende dataverzameling in het basisonderwijs, normen te kunnen ontwikkelen voor de groep 11-jarigen. We hebben ervoor gekozen normen voor 11-jarigen te ontwikkelen omdat een klein percentage leerlingen in leerjaar 1 (ongeveer 1 procent halverwege maart van het schooljaar) nog in deze leeftijdsgroep valt.
In overeenstemming met de oorspronkelijke opzet zijn er daarnaast normen ontwikkeld voor leerjaar 1, 2 en 3 van het voortgezet onderwijs.

4.3.1 Normconstructie op basis van weging

Achtergrondvariabelen
Voor de steekproef beschikten we over de volgende gegevens: leeftijd (afgeleid van geboortedatum en afnamedatum), leerjaar, onderwijstype, sekse, thuistaal en regio. Van deze variabelen zijn alleen de eerste drie gebruikt om te wegen. De overige variabelen werden niet in de weging betrokken.
De variabele sekse is niet in de weging betrokken omdat we ervan uit mogen gaan dat jongens en meisjes volgens de populatieverdeling in de steekproef gerepresenteerd zijn. Er zijn immers in de regel volledige schoolklassen bij de afnames betrokken geweest. Overigens zullen we verderop in hoofdstuk 6 laten zien dat sekse er niet toe doet bij het maken van de test: jongens en meisjes scoren bij benadering gemiddeld even hoog.
Voor regio (en verstedelijking; in het basisonderwijs ook voor stratum) geldt dat de scholen zijn geworven in overeenstemming met de populatieverdeling. Waar in de feitelijke deelname aan het onderzoek hiaten in de steekproef van scholen ontstonden, zijn deze hiaten door middel van aanvullend onderzoek zo veel mogelijk gedicht. Afgezien daarvan is het niet te verwachten dat regio er (op leerlingniveau) veel toe doet (zie handleiding NIO, van Dijk & Tellegen, 2004).
De variabele thuistaal lijkt daarentegen wel effect te hebben op de score (zie hoofdstuk 6 voor gegevens over de invloed van thuistaal op de testscores in de verschillende onderwijstypen). Helaas waren voor die variabele geen gedifferentieerde populatiegegevens te achterhalen, zodat weging in overeenstemming met de populatieverdeling niet mogelijk was. Op basis van de gehanteerde wervingsprocedures en de verwachte samenhang van thuistaal met de achtergrondvariabelen regio, verstedelijking (en in het basisonderwijs stratum), nemen we aan dat de steekproef naar thuistaal representatief is.

Gegevens over de populatieverdelingen
Er zijn geen gedifffferentieerde populatiegegevens bekend over de exacte verdeling van school- en onderwijstypen per leeftijd en leerjaar, Wel zijn er deelgegevens beschikbaar die het mogelijk maken deze verdeling zelf te genereren, waarbij we consistentie met de ons bekende gegevens hebben nagestreefd. Dit is gebeurd in een aantal stappen die we in het navolgende zullen beschrijven.

We hebben een aantal betrouwbare bronnen, zoals CFI-DUO en CBS, geraadpleegd met de bedoeling om voor iedere cel in de drie-wegtabel van leeftijd bij schooltype bij leerjaar, de populatieproporties te verkrijgen. De belangrijkste bron was het *Centraal Bureau van de Statistiek (CBS)*. Ook de gegevens van CFI-DUO waren bruikbaar, maar een vrij grote groep leerlingen in de leeftijd 12, 13 en 14 jaar is in de CFI-DUO-bestanden opgenomen in de categorie “brugjaar algemeen”. Daardoor is niet direct inzichtelijk hoe de leerlingen in deze leeftijdsgroepen over de verschillende onderwijsniveaus verdeeld zijn (zie tabel 4.5).

Het CBS biedt zicht op het aantal leerlingen per onderwijsniveau in de leerjaren 1, 2 en 3 van het voortgezet onderwijs (zie tabel 4.6). Door deze informatie te combineren met de informatie uit tabel 4.4 waarin per leeftijdsjaar de verdeling van leerlingen over de verschillende jaargroepen van het (speciaal) basis- en voortgezet onderwijs is gegeven (per 15 maart van het schooljaar), is de populatieverdeling geschat die in tabel 4.7 is weergegeven. Daarbij komen de marginalen zoveel mogelijk overeen met de gegevens van de afzonderlijke tabellen 4.4 en 4.6. We geven nu wat gedetailleerder weer in welke stappen we te werk zijn gegaan. Om te beginnen beschrijven we hoe tabel 4.3 tot stand is gekomen waarin we de leeftijdverdeling per leerjaar op 31 december weergeven. En vervolgens geven we aan hoe deze verdeling is doorgerekend naar de situatie op 15 maart van datzelfde leerjaar (tabel 4.4).
In tabel 4.3 staat vetgedrukt de aan het CBS ontleende verdeling van leeftijden 10 tot en met 15 jaar over de leerjaargroepen in het voortgezet onderwijs (leerjaar 1 tot en met 5) in het schooljaar 2008-2009 (in absolute aantallen leerlingen). Soortgelijke gegevens voor het basisonderwijs zijn er echter niet: we kennen voor het basisonderwijs alleen de totalen per leeftijd, niet de verdeling van leeftijden over leerjaren. De tabel is aangevuld met de totale aantallen leerlingen uit het basis- en speciaal basisonderwijs per leeftijdsgroep (voor schooljaar 2008-2009). De verdeling van leeftijden over de verschillende jaargroepen in het basisonderwijs hebben we moeten schatten. We deden dit op basis van de globale verdeling van leeftijden over de eerste drie leerjaren binnen het voortgezet onderwijs, onder de aanname dat deze verdeling voor

leerjaar 6, 7 en 8 van het basisonderwijs ongeveer vergelijkbaar is. De geschatte verdeling is in de tabel cursief weergegeven.

Tabel 4.3 Populatieverdeling naar leeftijd en leerjaar per 31 december (toelichting in de tekst)

	Leerjaar								
Leeftijd	6	7	8	1	2	3	4	5	
10	*77000*	*105000*	*2000*	**40**	**1**				184041
11	*7000*	*75000*	*103000*	**2139**	**42**	**7**	**2**	**1**	187191
12	*600*	*7000*	*71000*	**105337**	**1915**	**51**	**3**	**0**	185906
13		*1000*	*10000*	**75858**	**103875**	**1808**	**47**	**3**	192591
14			*400*	**7510**	**80599**	**102341**	**1595**	**37**	192482
15				**785**	**9611**	**83428**	**95644**	**1241**	190709

Een overzicht van het totale aantal leerlingen in 2008/2009 in het basisonderwijs en speciaal basisonderwijs volgens het CBS is opgenomen in de bijlagen (tabel B3). De totalen van de leerjaren 6,7 en 8 in tabel 4.3 corresponderen voor de 10-jarigen niet met de totalen van de CBS tabel die in de bijlage is opgenomen (tabel B.3) omdat een groter deel van de leerlingen nog in leerjaar 4 en 5 zit (circa 9000) en deze zijn in tabel 4.3 niet weergegeven. Voor de 11 tot en met de 14 jarigen is het verschil met de totalen van CBS te verwaarlozen (maximaal 200 leerlingen per leeftijdsgroep).

Tabel 4.4 Populatieverdeling naar leeftijd en leerjaar per 15 maart (toelichting in de tekst)

	Leerjaar								
Leeftijd	6	7	8	1	2	3	4	5	
11	21583	81250	81958	1702	33	6	2	1	186535
12	1933	21167	77667	83837	1525	42	3	0	186174
13	125	2250	22708	81999	82633	1442	38	2	191198
14	0	208	2400	21749	85448	81397	1273	30	192505
15	0	0	83	2186	24400	87368	76050	990	191078

	Leerjaar								
Leeftijd	6	7	8	1	2	3	4	5	
11	11,6	43,6	43,9	0,9	0,0	0,0	0,0	0,0	100
12	1,0	11,4	41,7	45,0	0,8	0,0	0,0	0,0	100
13	0,1	1,2	11,9	42,9	43,2	0,8	0,0	0,0	100
14	0,0	0,1	1,2	11,3	44,4	42,3	0,7	0,0	100
15	0,0	0,0	0,0	1,1	12,8	45,7	39,8	0,5	100

Tabel 4.5 Verdeling van onderwijsniveaus over leeftijdsgroepen op basis van de gegevens van CFI-DUO omgerekend naar het gemiddelde afnamemoment van het normeringsonderzoek

Leeftijd op 15 maart		*11*	*12*	*13*	*14*	*15*	*16*
BO	Stratum 1	58,1	31,8	4,9	0,1	0,0	0,0
	Stratum 2	21,4	12,3	2,4	0,1	0,0	0,0
	Stratum 3	10,7	6,9	1,9	0,1	0,0	0,0
	sbo	4,9	4,4	2,1	0,3	0,0	0,0
	so	2,7	2,1	1,1	0,5	0,4	0,4
VO	brugklas	1,0	22,3	32,8	16,3	3,7	0,5
	pro	0,0	0,2	0,8	1,1	1,1	1,2
	vmbo	0,2	12,6	33,5	47,0	52,7	42,8
	havo	0,0	1,1	6,0	13,7	18,8	27,5
	vwo	0,6	5,5	12,9	18,5	20,5	24,2
	es	0,1	0,1	0,1	0,2	0,2	0,1
	ib	0,0	0,0	0,0	0,0	0,0	0,1
	vso	0,1	0,6	1,5	2,2	2,6	3,2
		100,0	100,0	100,0	100,0	100,0	100,0
BO		90,3	51,0	9,2	0,3	0,0	0,0
SBO		7,7	6,5	3,2	0,7	0,4	0,4
VO		2,0	42,5	87,6	99,0	99,6	99,6

De verdeling naar leeftijd is in de CBS-tabellen standaard gebaseerd op 31 december van het betreffende schooljaar. Omdat de dataverzameling voor het normeringsonderzoek (zowel het initiële onderzoek in 2009 als het aanvullende onderzoek in 2010) heeft plaatsgevonden op of rond (gemiddeld) 15 maart was het zaak de verdeling van leeftijden over leerjaren op peildatum 15 maart te achterhalen. De berekening hiervoor is opgevraagd bij het CBS (15 maart zit op 5/24 van het jaar).
Het aantal 11-jarigen in groep 7 op 15 maart is dan (bij wijze van illustratie) af te leiden uit (5/24 * aantal 10-jarigen in groep 7 op 31 december) + (19/24 * aantal 11-jarigen in groep 7 op 31 december). Voor alle cellen van de tabel is deze berekening gemaakt; de aantallen zijn weergegeven in Tabel 4.4 voor de 11- tot en met 15-jarigen. Uiteindelijk zijn de absolute aantallen omgezet naar percentages per leeftijdsjaar.

Tabel 4.6 Verdeling van leerlingen over onderwijsniveaus in leerjaar 1,2 en 3 (CBS Statline 2008/2009; gegevens exclusief speciaal onderwijs)

	Leerjaar		
Onderwijsniveau (schooltype)	*1*	*2*	*3*
vwo	10,1	17,3	21,8
havo/vwo	22	12,4	
havo	2,5	11,7	20,2
vmbo-gt/havo/vwo	10	3,9	3,4
vmbo-gt/havo	10,6	4,6	
vmbo-gt	10,1	16	26,1
vmbo/havo	5,9	4,7	
vmbo	10,7	10,3	
vmbo bk	8,9	8,9	
vmbo-k	3,7	4,4	15,0
vmbo-b	5,6	5,9	13,5

In de volgende stappen was het nodig om de op deze manier afgeleide verdeling naar leeftijd en leerjaar aan te vullen met gegevens over het onderwijsniveau. Zoals eerder aangegeven was er helaas geen bron beschikbaar die de populatieverdeling voor de drieweg leeftijd-leerjaar-onderwijsniveau direct kon verstrekken. Wél waren beperkte CFI-DUO-gegevens beschikbaar over de verdeling van onderwijsniveaus over leeftijdsgroepen (beperkt vanwege de brede categorisering 'brugjaren algemeen'voor de leeftijdsgroepen 12, 13 en 14 jaar; zie tabel 4.5). Via CBS-statline is de verdeling van leerlingen in leerjaar 1, 2 en 3 over de verschillende onderwijsniveaus bekend; de gegevens zijn weergegeven in tabel 4.6.

Door de gegevens in de tabellen 4.4 en 4.6 te combineren is een goede indicatie van de populatieverdeling van onderwijsniveau naar leerjaar per leeftijdsgroep tot stand gekomen. Tabel 4.7 geeft deze verdeling weer in de vorm van proporties kinderen in de populatie voor elke leeftijd afzonderlijk, voor elke combinatie van onderwijsniveau (schooltype) en leerjaar. Voor elke leeftijd wil hier zeggen dat de proporties in elke subtabel per leeftijdsjaar optellen tot 1. Het zijn deze afgeleide en deels geschatte populatieproporties die gebruikt zijn voor de weging. Dat weging noodzakelijk was, blijkt uit het feit dat proporties in de gerealiseerde steekproef in sommige cellen afweken van de populatieproporties. In de tabel 4.7 is aangegeven waar dit met name het geval was. De cijfers zijn vetgedrukt wanneer het percentage in de steekproef meer dan 10% lager was dan het corresponderende percentage in de populatie. De percentages zijn schuingedrukt wanneer dit percentage meer dan 10% hoger was. De exacte verdeling over onderwijsniveau, leerjaar en leeftijd in de steekproef wordt in de volgende sectie besproken.

Tabel 4.7 Geschatte populatieverdeling naar leeftijd, leerjaar en onderwijstype (in proporties; toelichting in tekst)

Onderwijstype per leeftijd	*Basisonderwijs leerjaar*			*Voortgezet onderwijs leerjaar*				*Marginalen*
11 jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO	0,11	*0,42*	**0,42**					0,95
SO	0,01	0,02	0,02					0,05
vmbo								
havo								
vwo								
	0,12	0,44	0,44					
12 jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO	0,01	0,11	**0,41**					0,53
SO/VSO		0,01	0,02	0,02				0,05
vmbo				0,19				0,19
havo				*0,11*				0,11
vwo				0,11				0,11
	0,01	0,12	0,43	0,43				
13 jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO		0,010	0,120					0,13
SO/VSO		0,001	0,005	0,020	0,015			0,04
vmbo				0,230	0,230	0,002		0,46
havo				0,090	0,090	0,005		0,19
vwo				0,080	0,090	0,013	0,001	0,18
		0,011	0,125	0,42	0,425	0,02	0,001	
14 jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO			0,02					0,02
SO/VSO				0,02	0,02	0,01		0,05
vmbo				*0,05*	0,21	0,19		0,45
havo				0,02	0,11	0,10		0,23
vwo				0,01	0,10	0,13	0,01	0,25
			0,02	0,10	0,44	0,43	0,01	
15 jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO								
SO/VSO					0,01	0,02	0,01	0,04
vmbo				0,01	0,08	0,25	**0,18**	0,52
havo					0,03	0,09	**0,10**	0,22
vwo					0,03	0,09	0,12	0,24
				0,01	0,15	0,45	0,41	

Steekproefgegevens

Tabel 4.8 is qua opbouw gelijk aan tabel 4.7, maar bevat de proporties en aantallen leerlingen in *de steekproef in zijn oorspronkelijke samenstelling*.

Tabel 4.8 Oorspronkelijke steekproefproporties met absolute aantallen tussen haakjes

	Basisonderwijs leerjaar			*Voortgezet onderwijs leerjaar*				*Marginalen*
11 Jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO	0,08 (21)	0,59 (152)	0,26 (67)					0,94 (240)
SO		0,03 (6)	0,004 (1)					0,03 (7)
vmbo								
havo				0,008 (2)				0,008 (2)
vwo				0,03 (7)				0,03 (7)
	0,08 (21)	0,62 (158)	0,27 (68)	0,04 (9)				
12 Jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO		0,05 (28)	0,30 (167)					0,35 (195)
SO/VSO		0,03 (17)	0,04 (22)	0,002 (1)				0,07 (40)
vmbo				0,14 (80)				0,14 (80)
havo				0,26 (148)	0,002 (1)			0,27 (149)
vwo				0,16 (91)	0,01 (4)			0,17 (95)
		0,08 (45)	0,34 (189)	0,57 (320)	0,009 (5)			
13 Jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO		0,001 (1)	0,02 (15)					0,02 (16)
SO/VSO			0,02 (11)	0,01 (8)	0,01 (4)			0,03 (23)
vmbo				0,27 (191)	0,15 (106)			0,43 (297)
havo				0,19 (131)	0,11 (75)	0,001 (1)		0,30 (207)
vwo				0,14 (96)	0,08 (56)	0,003 (2)		0,22 (154)
		0,001 (1)	0,04 (26)	0,61 (426)	0,35 (241)	0,004 (3)		
14 Jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO								
SO/VSO				0,003 (2)	0,01 (6)	0,002 (1)		0,01 (9)
vmbo				0,20 (128)	0,28 (179)	0,09 (59)		0,56 (366)
havo				0,03 (18)	0,13 (84)	0,07 (42)		0,23 (144)
vwo				0,002 (1)	0,09 (55)	0,10 (61)		0,18 (117)
				0,23 (149)	0,51 (324)	0,27 (163)		
15 Jaar	*6*	*7*	*8*	*1*	*2*	*3*	*4*	
BO								
SO/VSO						0,02 (5)		0,02 (5)
vmbo				0,06 (17)	0,14 (43)	0,34 (103)	0,001 (1)	0,55 (164)
havo				0,003 (1)	0,02 (6)	0,16 (49)	0,01 (2)	0,19 (58)
vwo					0,02 (6)	0,15 (46)	0,07 (20)	0,24 (72)
				0,06 (18)	0,18 (55)	0,68 (203)	0,08 (23)	

Vergelijken we deze gegevens met de populatiegegevens dan is er geen statistische toets nodig om vast te stellen dat de proporties in een aantal cellen afwijken. De belangrijkste afwijkingen zijn, zoals eerder aangegeven, al gemarkeerd in tabel 4.7.

De wegingsfactor

Bij de toegepaste systematiek van steekproefweging is er vooral een probleem wanneer er percentueel *te weinig* kinderen van een bepaald type in de steekproef zitten. In dat geval bevat de gewogen steekproef *meer* kinderen van dit type dan de oorspronkelijke steekproef. De gewogen steekproeven worden immers met teruglegging getrokken en willekeurige kinderen worden herhaald opgenomen. De factor waarmee de steekproef groter wordt is het equivalent van de wegingsfactor.

Om te beoordelen of een wegingsfactor relevant is, moeten we rekening houden met de populatie-percentages. Om zicht te krijgen op de echte knelpunten, kunnen we ons beperken tot die gevallen waarin de populatie-percentages groter zijn dan 2%. Bij kleinere percentages is het effect van weging verwaarloosbaar.

In de volgende gevallen is sprake van een wegingsfactor groter dan 2.

- 12-jarigen in leerjaar 7 van het BO. De wegingsfactor is hier 2.2
- 13-jarigen in leerjaar 8 van het BO. De wegingsfactor is hier 5.58
- 14-jarigen in leerjaar 2 van het vmbo. De wegingsfactor is hier 2.05.
- 15-jarigen in leerjaar 4 van het vmbo. De wegingsfactor is hier 58.82
- 15-jarigen in leerjaar 4 van het vmbo. De wegingsfactor is hier 14.95.

In alle overige gevallen ligt de wegingsfactor onder de factor 2.
Op basis van deze uitkomsten is besloten geen normen te construeren voor 15-jarigen. Voor de overige leeftijdsgroepen was het aantal knelpunten beperkt.

4.3.2 Resultaten met betrekking tot de modelpassing

Kalibratie, DIF-analyses en R1c-toetsen
Er is een OPLM geschat voor elke van de zes deeltaken afzonderlijk. In eerste instantie werd elke deeltaak gekalibreerd op de data van alle leeftijden, dit onder aanname dat er geen DIF was ten aanzien van leeftijd. Deze aanname bleek echter niet houdbaar te zijn. Dat wil zeggen dat 'rijping' en onderwijs invloed hebben op de eigenschappen van de items. Dat wil zeggen, zelfs op de items van deze intelligentietest, waarbij we veronderstellen dat deze in niets (of althans zo weinig mogelijk) lijken op de taken die kinderen op school leren uitvoeren. Het is niet onwaarschijnlijk dat dit ook bij subtests van andere intelligentietests het geval is, zonder dat daarover veel bekend is. Bekende intelligentietests zoals de WISC-III zijn immers in de regel op klassieke wijze geconstrueerd waarbij gegevens over eventuele DIF naar leeftijd verborgen blijven.
De methodologie die is gebruikt om DIF te onderzoeken alsook een voorbeeld waarbij een subtest van een (andere) intelligentietest is gebruikt, is te vinden in Bechger en Maris (2010).

Omdat met een intelligentietest zoals de Intelligentietest VO geen uitspraken hoeven te worden gedaan over leeftijdsgroepen heen en evenmin over verschillen in vaardigheid (er wordt immers genormeerd in termen van een deviatie-IQ), konden we eenvoudigweg aparte kalibraties uitvoeren per leeftijdsgroep. Tabel 4.9 geeft de overschrijdingskansen van de R1c-toets voor elk van de per leeftijdsgroep en per deeltaak uitgevoerde kalibraties. De aantallen in de steekproef zijn te vinden in eerdere tabellen. Daar is te zien dat de 13-jarigen de grootste steekproef vormden met n = 697: het feit dat de schalen het slechtst lijken te passen bij de 13-jarigen is dus waarschijnlijk te wijten aan de grotere statistische power in die groep. In de tabel zijn alleen de vetgedrukte waarden significant nadat Bonferroni's correctie is uitgevoerd. Bij nadere beschouwing van de afzonderlijke items bleek dat hier een beperkt aantal items verantwoordelijk is voor de geconstateerde (relatief) mindere passing[8]. De conclusie mag luiden dat *overall* de R1c-toetsingen wijzen op een uitstekende passing van het meetmodel.

Tabel 4.9 R1c-toetsen met overschrijdingskansen voor de kalibraties van alle deeltaken per leeftijdsgroep

	R1c-toetsen: overschrijdingskansen per deeltaak					
Leeftijd	FC	FM	GR	GA	WC	WA
11	0,033	0,097	0,06	0,188	0,005	**0,002**
12	0,234	0,259	0,119	0,013	0,250	0,385
13	**0,003**	0,050	0,194	**0,0002**	**0,002**	0,093
14	0,404	0,019	**0,002**	0,422	0,020	0,129
15	0,18	0,60	0,01	0,08	0,005	0,73

[8] We tekenen hierbij aan dat de deeltaken aanvankelijk zijn samengesteld op basis van klassieke testtheorie. IRT-kalibratie is in tweede instantie uitgevoerd op de normeringsgegevens.

Discriminatie-indices en de efficiëntie van de ongewogen scores
Bij de meeste deeltaken vertoonde het Raschmodel een goede passing en zijn discriminatie-indices incidenteel aangepast op basis van de empirische itemresponse curves (voor voorbeelden zie figuur 4.2 en 4.3). Bij enkele schalen zijn bij de kalibratie de discriminatie-indices geschat. In alle gevallen geldt dat de verdeling van de R1c hierdoor strikt genomen niet geheel correct is omdat we kapitaliseren op kans. Dit is een gevolg van de gehanteerde procedure en in de praktijk helaas onvermijdelijk. In dit geval ligt onze prioriteit echter bij het optimaal beschrijven van de data, zodat we er zeker van zijn dat de belangrijkste resultaten goed zijn: de gerepliceerde scores.
Zoals hierboven uiteengezet baseren we uitspraken over de intelligentie (in termen van leerjaar- en leeftijd-IQ) op de ongewogen scores, dus het aantal correcte antwoorden. Dit impliceert dat we, ten opzichte van gewogen scores, efficiëntie verliezen. Dit verlies aan efficiëntie is echter zeer gering. Dit blijkt uit correlaties tussen de gewogen en de ongewogen scores bij alle deeltaken. Hoewel een OPLM is gebruikt voor de kalibraties, liggen deze bij alle deeltaken onveranderlijk boven 0,99.

4.4 Normering en verdelingskenmerken

In deze paragraaf gaan we in op de vraag hoe de test uiteindelijk is genormeerd en geven we in het kort aan hoe de stap van de in deze wetenschappelijke verantwoording gerapporteerde analyses naar het scoringsprogramma en het leerlingrapport is genomen (paragraaf 4.4.1). Vervolgens bespreken we in paragraaf 4.4.2 een aantal kenmerken van (de verdeling van) de testscores.

4.4.1 Normering: van analyses naar leerlingrapport

Het digitale scoringsprogramma
Input voor het scoringsprogramma is het door de leerling ingevulde antwoordblad waarop deze zijn of haar antwoord op een item heeft aangestreept. Bij de verwerking vindt een check plaats of zich onregelmatigheden bij het aanstrepen hebben voorgedaan. Daarnaast wordt gecontroleerd of de leerling voldoende opgaven van een domein heeft gemaakt. Wanneer het antwoordblad niet aan de criteria voldoet wordt het van verdere verwerking uitgesloten en worden geen verdere resultaten gerapporteerd.
De itemscores worden geconverteerd in goed-foutscores op basis van sleutels. Informatie over de sleutels is desgewenst verkrijgbaar bij Cito. De sleutels zijn tevens als bijlage opgenomen in de handleiding.
Zoals in het begin van paragraaf 4.2 is aangegeven werken we in de Intelligentietest VO met deviatie-IQ's op basis van de traditionele verdelingskenmerken (normaal verdeeld met gemiddelde 100 en standaard-deviatie 15). In eerste instantie worden scores bepaald door itemscores (aantal goed) bij elkaar op te tellen tot een ongewogen somscore. Dat gebeurt per domein (Getallen, Woorden en Figuren); daarnaast worden ook alle itemscores bij elkaar opgeteld tot de totale testscore. De ruwe somscores worden niet gerapporteerd. Ook wordt niet gerapporteerd op het niveau van de afzonderlijke deeltaken.
Vervolgens worden de IQ-scores bepaald. Er wordt een leeftijd- en een leerjaar-IQ vastgesteld op basis van de cumulatieve verdelingen voor de leeftijdsgroep en leerjaargroep waarin de leerling valt (i.e. in de naar leerjaar, leeftijd en schooltype gewogen steekproef). Deze zijn in het scoringsprogramma ondergebracht in de vorm van normtabellen. Daarbij wordt een zogeheten leerdagcorrectie toegepast (een vorm van interpolatie; meer hierover in de volgende sectie). Naast het leerjaar- en leeftijd-IQ worden ook de bijbehorende percentielscores afgedrukt op het leerlingrapport en grafisch weergegeven. Ook voor de domeinscores (Getallen, Woorden en Figuren) worden percentielscores gerapporteerd. Op basis van informatie over de lokale meetfout wordt de rapportage van zeer lage en zeer hoge IQ's gefixeerd op waarden van 70, respectievelijk 130.
Belangrijker nog dan de IQ-scores zelf zijn de betrouwbaarheidsintervallen. Op het leerlingrapport zijn voor de beide IQ-scores de 80%-betrouwbaarheidsintervallen weergegeven. Deze betrouwbaarheidsintervallen zijn bepaald volgens de procedures die we in paragraaf 4.2.4 hebben beschreven. Elke IQ-score kent dus zijn eigen betrouwbaarheidsinterval.
Ten slotte wordt het leerjaar-IQ grafisch weergegeven in relatie tot de verdelingskenmerken per schooltype (vwo, havo, vmbo-gt, vmbo-kb, vmbo-bb met en zonder lwoo). Daartoe is op het leerlingrapport

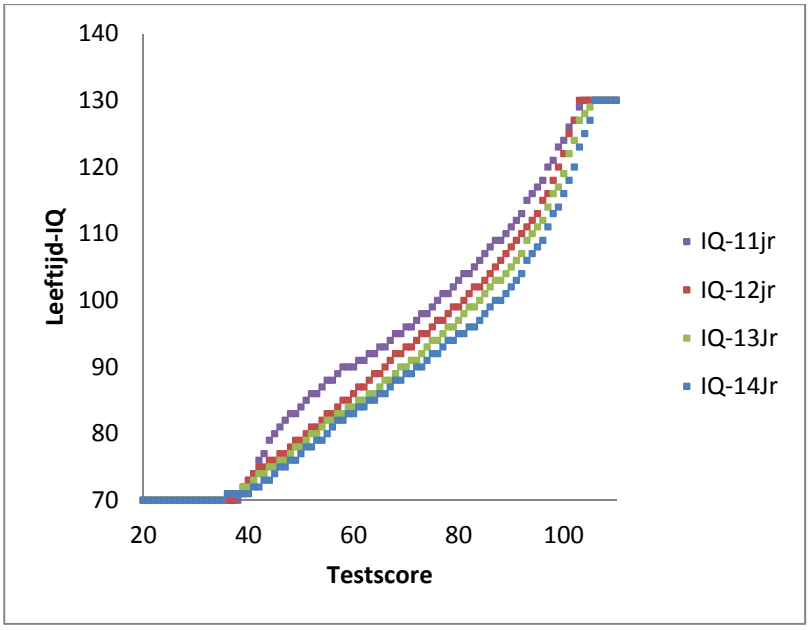
aangegeven waar de mediane leerjaar-IQ-score per schooltype ligt en per schooltype de IQ-score die hoort bij de 20% laagst- en hoogstscorenden. Gegevens hierover zijn te vinden in hoofdstuk 7.

Leerdagcorrectie
Zoals eerder opgemerkt is het – ook in dit instrument gehanteerde – deviatie-IQ een getal dat uitdrukt hoe een score zich verhoudt tot een referentieverdeling. Traditioneel bestaat de referentiegroep uit leeftijd-genoten waarbij op de leeftijd van 11 tot en met 14 jaar het gebruik van leeftijdgroepen die gedefinieerd zijn in termen van gehele jaren geen problemen oplevert; in figuur 4.3 is te zien hoe dicht de verdelingen voor de onderscheiden leeftijdsgroepen tegen elkaar aanliggen. Hoe nauwkeurig de test ook meet (zie hoofdstuk 5 voor de betrouwbaarheden), de betrouwbaarheidsintervallen voor de 12- en 14-jarigen blijken elkaar grotendeels te overlappen. In principe kan een referentiegroep bovendien om praktische redenen gedefinieerd worden door elk willekeurig criterium. Wij gebruiken hier naast leeftijdsgroepen ook leerjaargroepen als referentiekader.

Wanneer een kind wordt vergeleken met leeftijdgenoten met dezelfde kalenderleeftijd, impliceert dit dat, wanneer een kind jarig is, de referentiegroep van de ene op de andere dag verandert en daarmee het IQ. Methodologisch gezien is hier geen enkel bezwaar tegen omdat een kind dan wordt vergeleken met een andere referentiegroep. Gebruikers lijken soms moeite mee te hebben met dit verschijnsel. Dit is begrijpelijk omdat de vaardigheidsscore die de basis vormt van het IQ (in theorie) van de ene dag op de andere nauwelijks verandert. Men lijkt daarom bij voorkeur een IQ te willen kunnen bepalen waarbij de referentie-groep bestaat uit kinderen die op de maand af even oud zijn. Men dwingt daarmee de testontwikkelaar als het ware om een vorm van interpolatie of continue normeren te gebruiken.
In vrijwel alle bekende toepassingen is continue normeren gebaseerd op een vorm van *kwantiel-regressie* (Koenker, 2005). In de huidige context houdt kwantielregressie in dat we de relatie modelleren tussen kwantielen van de scoreverdeling en de leeftijd. Bij continue normeren wordt deze relatie vervolgens gebruikt om de kwantielen (en daarmee het IQ) te schatten in veel fijnmaziger gedefinieerde normgroepen. Een voorbeeld moge dit toelichten. Hoewel er weinig kinderen in de steekproef zijn van exact 12 jaar en 2 maanden wordt de regressiefunctie gebruikt om de verdeling van deze groep te schatten. In de praktijk is de regressiefunctie doorgaans een hoge-orde polynoom, die in feite meer bedoeld is om zo goed mogelijk de data te beschrijven (“curve-fitting”) dan om te modelleren.
Het gebruik van continue normeren heeft, in het algemeen, voordelen voor een normering aangezien informatie van alle normgroepen simultaan wordt gebruikt om de normen voor alle groepen te schatten. Naar onze mening kennen we de statistische eigenschappen van kwantielregressie echter nog onvoldoende. Hoewel we van plan zijn om de techniek na nader onderzoek te gebruiken in toekomstige uitgaven van deze intelligentietest, hebben we er bij het vaststellen van de huidige normen geen gebruik van gemaakt.

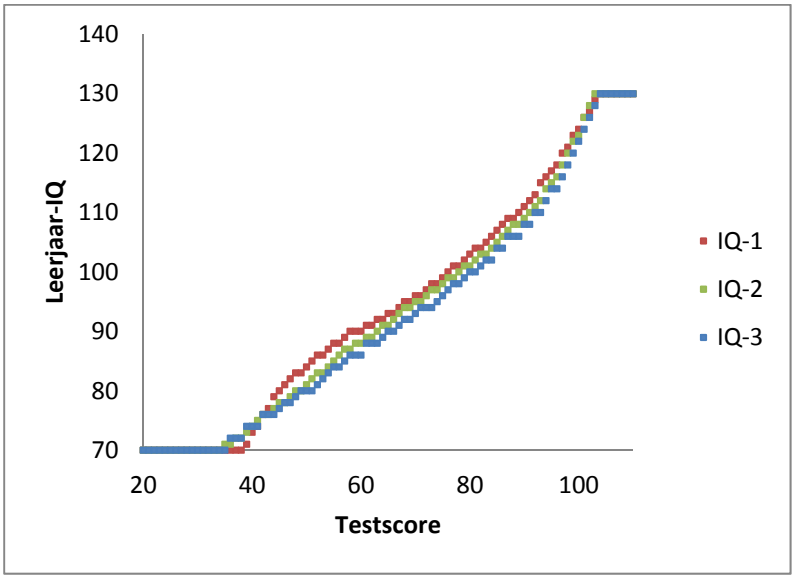
Om toch aan de wensen van het veld tegemoet te komen hebben we de IQ-score tabellen simpelweg geïnterpoleerd. Hoe dat in zijn werk gaat is goed te illustreren aan de hand van figuur 4.6, waarin de relatie tussen de testscore en de (genormeerde) IQ-score is weergegeven voor verschillende leeftijdsgroepen. Stel, we beschouwen iemand die zojuist 12 jaar geworden is. Die krijgt dan, gegeven een bepaalde testscore, een IQ dat ligt tussen het IQ voor de 11-jarigen en dat voor de 12-jarigen. Betrouwbaarheids-intervallen worden op analoge wijze aangepast. Het maximale verschil bij overstap van de ene in de andere normtabel met leerdagcorrectie is voor het leeftijd-IQ nooit groter dan 1 IQ-punt in plaats van maximaal 4 IQ-punten zonder de leerdagcorrectie. Dit kon gerealiseerd worden door de normtabellen met leerdagcorrectie per maand nauwkeurig op te stellen.

Figuur 4.6 De relatie tussen leeftijd-IQ en testscore



Ditzelfde is gedaan voor de leerjaar-IQ's. Figuur 4.7 toont de afbeelding van leerjaar-IQ op scores voor de verschillende leerjaren. De ordening tussen leerjaren is zoals verwacht: kinderen in hogere leerjaren krijgen nooit een hoger IQ bij dezelfde testscore. Ook bij de vaststelling van het leerjaar-IQ leidt de doorgevoerde leerdagcorrectie tot kleine verschillen in scores bij de overgang van de ene normtabel naar de andere (nooit meer dan 2 IQ-punten in plaats van maximaal 5 IQ-punten zonder leerdag correctie). Dit is gerealiseerd door normtabellen op te stellen per kwartaal.

Figuur 4.7 Relatie tussen leerjaar-IQ en testscore.



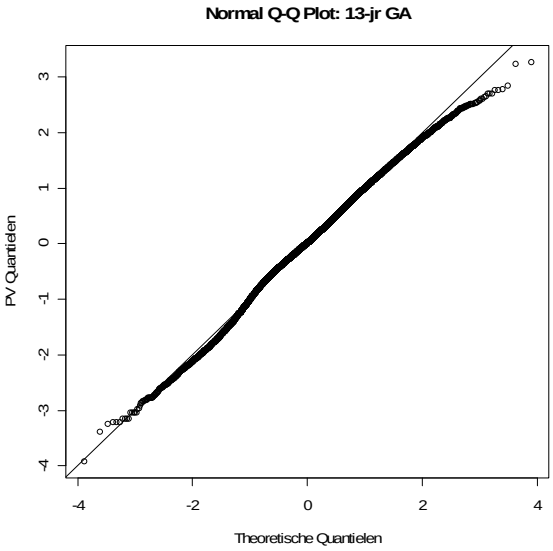
Bij figuur 4.6 en 4.7 valt nog op te merken dat extreme IQ-waarden van lager dan 70 en hoger dan 130 niet worden verstrekt. De test lijkt immers weinig geschikt om over deze scoreregionen betrouwbare uitspraken te doen (zoals waarschijnlijk in sommige andere intelligentietests ook het geval is, zonder dat de rapportage

daar doorgaans op is aangepast). Er zitten immers te weinig heel gemakkelijke of heel moeilijke items in de test en de scoreverdelingen worden in de staarten niet optimaal geschat (vergelijk figuur 4.8. verderop in dit hoofdstuk).

4.4.2 Verdelingskenmerken

Assumpties
We veronderstellen *a priori* dat de vaardigheden die met deeltaken van deze intelligentietest worden gemeten multivariaat normaal verdeeld zijn. Het gebruik van PVs laat echter toe dat de *geschatte* verdeling dat niet is. De marginale verdeling van de PVs is een consistente schatter van de ware vaardigheids-verdeling en daaraan kunnen we dus zien of de aanname van normaliteit houdt of niet.
Na weging vonden we, gebruikmakend van de toets van Jarque en Bera (1987), een significante afwijking van multivariate normaliteit. De marginale verdelingen laten echter geen of slechts kleine afwijkingen van normaliteit zien. De grootste afwijking vonden we bij Getallen. Ter illustratie toont figuur 4.8 een plot van kwantielen van de (gestandaardiseerde) PV-verdeling van de vaardigheid GA voor 13-jarigen tegen de kwantielen van de standaardnormale verdeling. Afwijkingen van de eerste bissectrice duiden op afwijkingen van normaliteit. Het beeld in de figuur geeft aan dat de verdeling iets gepiekter is dan de normale en niet helemaal symmetrisch maar enigszins scheef met meer subjecten in de lage staart van de verdeling dan op basis van de normaalverdeling verwacht mocht worden.

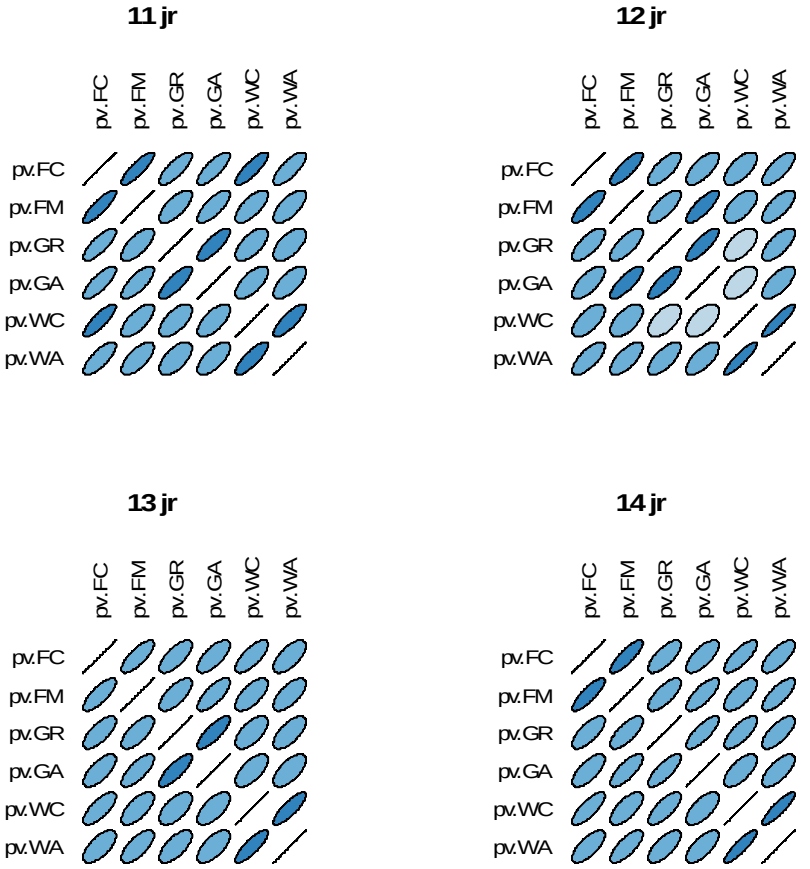
Figuur 4.8 QQ-plot voor de vaardigheidsverdeling van de deeltaak GA (13-jarigen)



Een andere assumptie betreft de samenhang tussen de vaardigheden die met verschillende deeltaken worden gemeten. We verwachten hoge correlaties tussen deze vaardigheden zodat het verantwoord is om de verschillende deeltaakscores in één testscore (als basis voor leerjaar- en leeftijd-IQ) te combineren. Daarbij moeten we rekening houden met het feit dat er DIF geconstateerd is tussen de onderscheiden leeftijdsgroepen en dat de vaardigheden die we meten voor verschillende leeftijdsgroepen dus kwalitatief wat anders en dus niet helemaal vergelijkbaar zijn. Correlaties tussen vaardigheden voor specifieke leeftijdsgroepen kunnen wél zinnig worden geïnterpreteerd.
Zoals verwacht zien we voor alle leeftijdsgroepen afzonderlijk hoge tot zeer hoge positieve correlaties tussen de vaardigheden gemeten met de deeltaken. De correlaties tussen de PVs voor de vaardigheden die met de deeltaken worden gemeten liggen in het bereik 0,58 – 0,90. Een overzicht van alle intercorrelaties per leeftijd- en leerjaargroep is te vinden in de bijlagen. Deze intercorrelaties laten voor de onderscheiden normgroepen allemaal hetzelfde patroon zien. We proberen dit (grafisch) te illustreren voor

de vier leeftijdsgroepen (zie figuur 4.9). Ieder element in de matrix is vervangen door de ellips van een bivariate normale verdeling waarvan de vorm en de kleur corresponderen met de betreffende correlatie (zie Murdoch en Chow, 1996). Te zien is dat correlaties tussen deeltaken die te maken hebben met G (getallen), F (figuren), of W (woorden) onderling hoger zijn dan de correlaties tussen deeltaken uit verschillende domeinen.

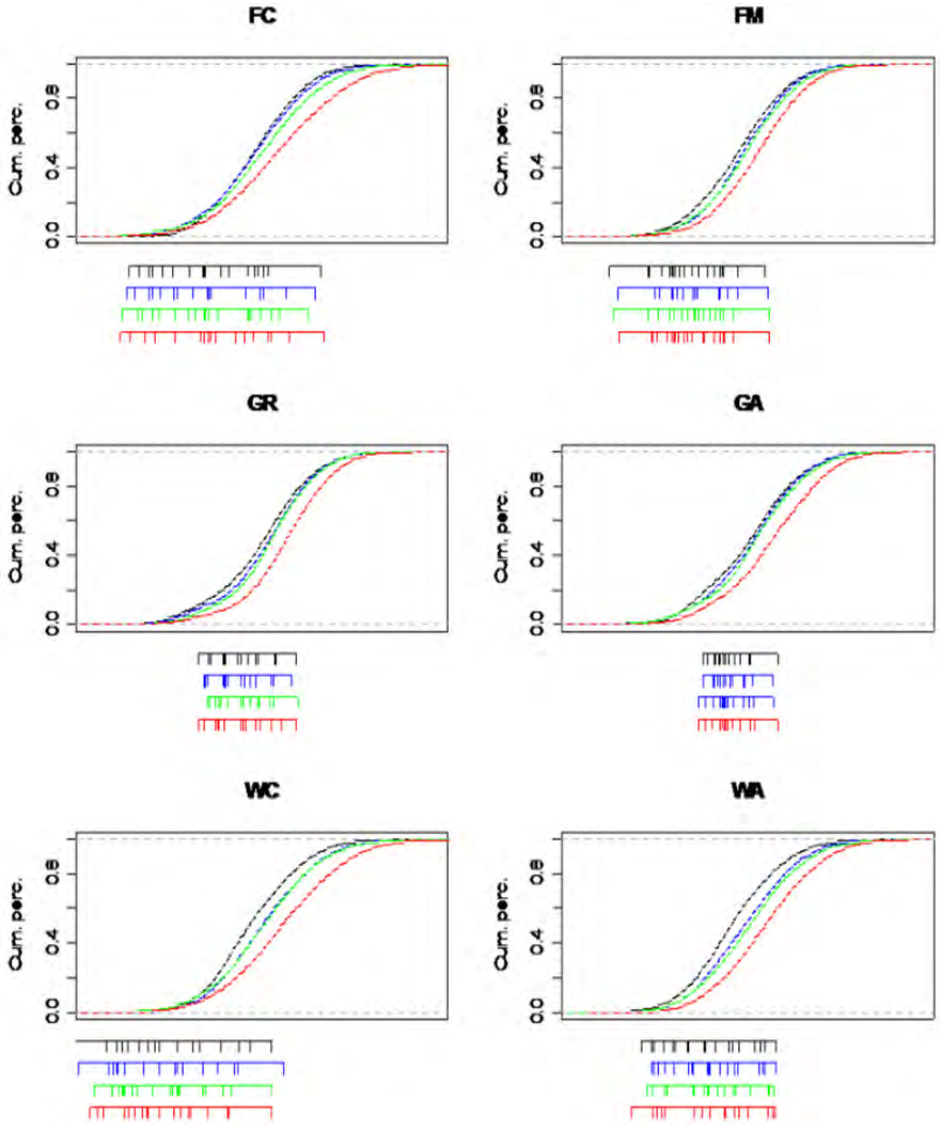
Figuur 4.9 Grafische representatie van de correlatiematrix tussen deeltaakvaardigheden voor de vier onderscheiden leeftijdgroepen (hoe hoger de correlatie, des te 'blauwer' en smaller de ellips)



De bijbehorende getallen (zie bijlagen) laten bijvoorbeeld voor de 12-jarigen zien dat de gemiddelde intercorrelatie 0,72 bedraagt. De correlaties tussen FC en FM (0,82), tussen GR en GA (0,83) en tussen WC en WA (0,90) zijn aanzienlijk hoger dan dit gemiddelde. De correlaties laten dus zien dat het verantwoord is om enerzijds somscores te berekenen per domein (Figuren, Woorden en Getallen) en anderzijds de deeltaakscores te combineren in één testscore. In hoofdstuk 6 Validiteit wordt hier uitgebreider op ingegaan.

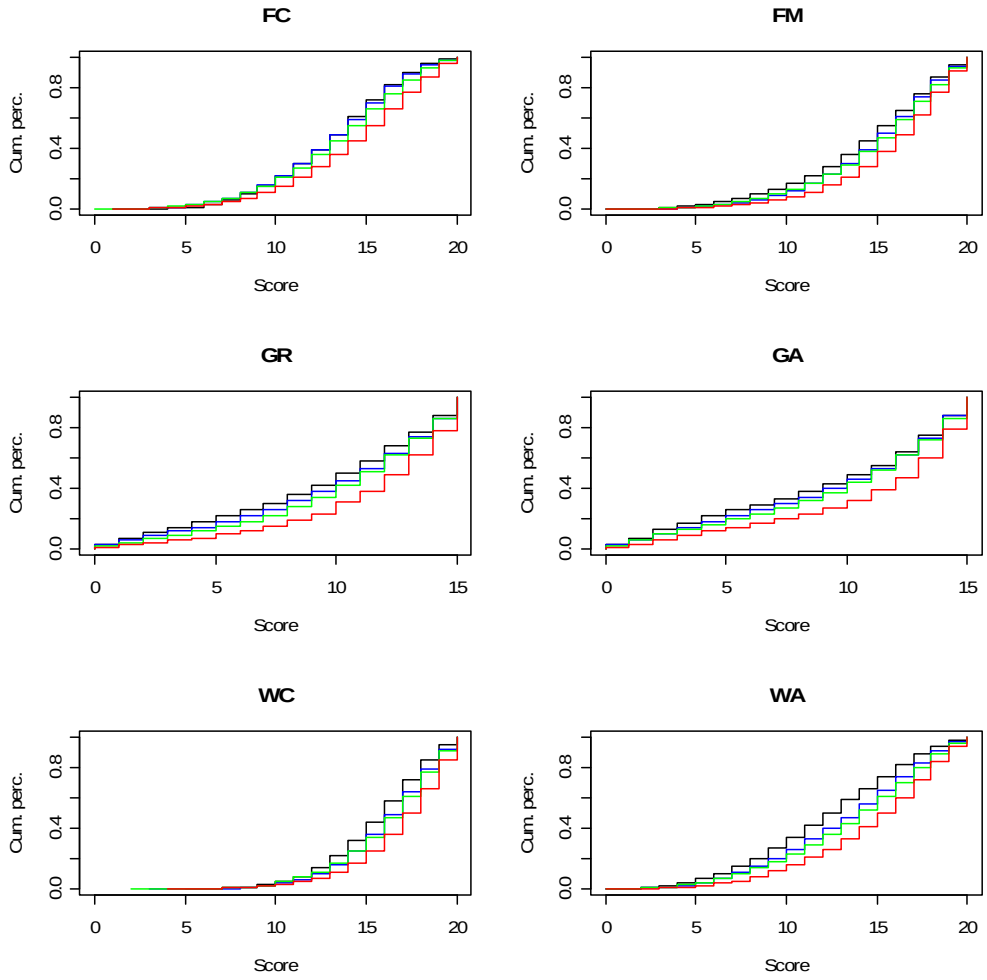
Vaardigheidsverdelingen voor verschillende leeftijden in relatie tot itemmoeilijkheid
In figuur 4.10 is voor elke deelvaardigheid, de verdeling getekend voor vier leeftijdsgroepen: 11, 12, 13 en 14 jarigen. We kunnen de vaardigheidsverdelingen niet direct onderling vergelijken maar alleen in relatie tot de items. De assen onder ieder figuur geven de moeilijkheden van de items weer als locaties op de vaardigheid-as.

Figuur 4.10 Geschatte vaardigheidsverdelingen voor 11- (zwart), 12- (blauw), 13- (groen) en 14-jarigen (rood) Itemmoeilijkheden in de leeftijdgroepen staan op additionele assen onderaan elke figuur.



Voor een vergelijking tussen leeftijden (en ook tussen schooltypen; zie daarvoor hoofdstuk 6 over validiteit) kunnen we de vaardigheidsverdelingen die in figuur 4.10 zijn weergegeven niet gebruiken, maar wel de scoreverdelingen. Preciezer uitgedrukt, de verdelingen van de gerepliceerde scores. Deze verdelingen zijn gebaseerd op de gewogen vaardigheidsverdelingen. Het effect van het gebruik van gerepliceerde scores in plaats van geobserveerde scores is dat de verdelingen zijn “gladgestreken”. Dit leidt in geen geval tot verschillende conclusies over de aard van en de verhoudingen tussen de vaardigheidsverdelingen, hetgeen nadere evidentie oplevert dat het IRT-model past bij de data.
In figuur 4.11 zijn de scoreverdelingen op alle zes deeltaken voor vier groepen van verschillende leeftijd afgebeeld. De verdelingen zijn geordend zoals men mag verwachten, de oudere kinderen tenderen hoger te scoren. Met name tussen de 11-, 12- en 13-jarigen zijn de verschillen echter klein. Ook opvallend is dat de mediane score vrijwel altijd in de hogere scores ligt. Zoals we verderop in hoofdstuk 6 zullen zien gaat dit niet meer op wanneer we de groepen uitsplitsen naar schoolniveau.

Figuur 4.11 Scoreverdelingen voor verschillende leeftijdsgroepen op de zes deeltaken



Ten slotte geven we in tabel 4.10 de kenmerken (gemiddelde, SD, skewness en kurtosis) weer voor alle scores die uiteindelijk in de vorm van normscores in het leerlingrapport worden gerapporteerd, dat wil zeggen de ruwe somscore voor Figuren, Getallen en Woorden, en de totale testscore. We doen dit voor elke leeftijdsgroep afzonderlijk. Voor de volledigheid geven we in tabel 4.11 ook de kenmerken van de scoreverdelingen voor de drie leerjaargroepen. En tabel 4.12 geeft de moeilijkheid weer in de verschillende normgroepen voor de totaalscore en de score op de domeinen.

De gemiddelden lopen met de leeftijd op, waarbij de standaarddeviaties per leeftijdsgroep ongeveer aan elkaar gelijk zijn. De spreiding voor Getallen is wat groter dan voor Figuren en Woorden. Voor de verdelingskenmerken voor de drie leerjaargroepen kunnen dezelfde conclusies worden getrokken. Gezien de grote mate van parallelliteit tussen leeftijd en leerjaar is dat ook wel te verwachten.

Tabel 4.10 Gemiddelde, SD, skewness en kurtosis voor de vier leeftijdsgroepen

11-jaar

	Gem.	SD	Skewness	Kurtosis
F	27.56	6.60	-0.55	2.75
G	18.84	8.44	-0.56	2.12
W	27.82	6.15	-0.32	2.51
Totaal	74.22	18.82	-0.42	2.25

12 jaar

	Gem.	SD	Skewness	Kurtosis
F	28.11	6.40	-0.65	3.06
G	19.73	8.22	-0.72	2.41
W	29.40	6.20	-0.52	2.65
Totaal	77.23	18.34	-0.58	2.60

13 jaar

	Gem.	SD	Skewness	Kurtosis
F	28.51	6.67	-0.72	3.14
G	20.30	7.79	-0.78	2.63
W	29.78	6.31	-0.65	2.85
Totaal	78.59	18.33	-0.72	2.90

14 jaar

	Gem.	SD	Skewness	Kurtosis
F	30.24	6.34	-0.85	3.52
G	22.49	7.24	-1.16	3.48
W	31.59	5.92	-0.87	3.29
Totaal	84.32	17.37	-0.92	3.29

Tabel 4.11 Gemiddelde, SD, skewness en kurtosis voor de vier leerjaargroepen

Leerjaar 1

	Gem.	SD	Skewness	Kurtosis
F	28.39	6.39	-0.68	3.14
G	20.19	7.92	-0.73	2.49
W	29.91	6.20	-0.63	2.79
Totaal	78.49	18.09	-0.65	2.71

Leerjaar 2

	Gem.	SD	Skewness	Kurtosis
F	29.41	6.49	-0.91	3.67
G	21.58	7.36	-0.99	3.12
W	30.86	6.27	-0.87	3.29
Totaal	81.86	17.82	-0.93	3.43

Leerjaar 3

	Gem.	SD	Skewness	Kurtosis
F	30.59	6.37	-0.93	3.54
G	22.86	7.09	-1.16	3.42
W	31.79	6.12	-0.92	3.34
Totaal	85.25	17.74	-0.99	3.28

Tabel 4.12 Moeilijkheid totaal en per subdomein per normgroep

Domeinen	11 jaar	12 jaar	13 jaar	14 jaar
Totaal	0.67	0.70	0.71	0.77
Figuren	0.69	0.70	0.71	0.76
Getallen	0.63	0.66	0.68	0.75
Woorden	0.70	0.74	0.74	0.79

Domeinen	Leerjaar 1	leerjaar 2	leerjaar 3
Totaal	0.71	0.74	0.78
Figuren	0.71	0.74	0.76
Getallen	0.67	0.72	0.76
Woorden	0.75	0.77	0.79

5 Betrouwbaarheid

In dit hoofdstuk beschrijven we de resultaten van de verschillende betrouwbaarheidsonderzoeken en –analyses die voor de Cito Intelligentietest VO zijn uitgevoerd. In paragraaf 5.1 rapporteren we de betrouwbaarheden die we konden bepalen op basis van de eerder in hoofdstuk 4 bescheven Bayesiaanse procedures. In paragraaf 5.2 gaan we in op de test-hertestbetrouwbaarheid. Ten slotte besteden we in paragraaf 5.3 aandacht aan de lokale meetnauwkeurigheid van de test.
We zullen steeds betrouwbaarheidscijfers verstrekken op drie verschillende niveaus. Op de eerste plaats is dat het niveau van de algemene intelligentie, zoals dat wordt gerapporteerd in termen van een (algemeen) leeftijd- en leerjaar-IQ. Aan de betrouwbaarheid op dit niveau mogen de hoogste eisen worden gesteld omdat de gerapporteerde waarden aanleiding kunnen geven tot belangrijke en soms onomkeerbare beslissingen. Daarnaast geven we de betrouwbaarheden op het niveau van de drie onderscheiden domeinen (Woorden, Figuren en Getallen). Op dit niveau moeten de waarden voldoen aan minder strikte eisen (overeenkomend met de eisen die de COTAN stelt aan tests die gebruikt worden om minder belangrijke beslissingen te onderbouwen; vergelijk Evers et al., 2010). Ten slotte geven we voor de volledigheid ook de betrouwbaarheden voor de afzonderlijke deeltaken, hoewel op dit gedetailleerde niveau niet wordt gerapporteerd in het leerlingrapport.

5.1 Betrouwbaarheid op basis van Bayesiaanse schattingen

In paragraaf 4.2.4 zijn we uitvoerig ingegaan op de wijze waarop we (Bayesiaanse) betrouwbaarheids-intervallen hebben bepaald. We hebben beschreven dat we dit deden op basis van de a posteriori voorspelde (‘posterior predictive’) verdeling van de testscores. De gerepliceerde scores kunnen we gebruiken om de betrouwbaarheid van de testscores op basis van klassieke testtheorie te bepalen. Formele details met betrekking tot de relatie tussen klassieke testtheorie en IRT zijn te vinden in Bechger, Maris, Verstralen en Béguin (2003).

In Tabel 5.1 geven we de betrouwbaarheden voor de vier onderscheiden leeftijdsgroepen en in tabel 5.2 voor de drie leerjaar-normgroepen. De betrouwbaarheid van de IQ’s is in elk van de leeftijdsgroepen 11 tot en met 14 jaar gelijk aan 0,95. Hetzelfde geldt voor de leerjaargroepen (eerste, tweede en derde leerjaar van het VO). Merk op dat deze betrouwbaarheidswaarden niet moeten worden geïnterpreteerd als ondergrens van de “ware” betrouwbaarheid (in klassieke termen), zoals *Cronbach’s alpha*. De getallen representeren de betrouwbaarheid onder het marginale multi-dimensionele OPLM, waarbij de populatieverdeling gewogen is.

Tabel 5.1 Betrouwbaarheden deeltaken, domeinscores en leeftijd-IQ

		Deeltaken						*Domeinen*			*Leeftijd-*
		FC	*FM*	*GR*	*GA*	*WC*	*WA*	*F*	*G*	*W*	*IQ*
Leeftijd	11	0,70	0,80	0,91	0,91	0,61	0,81	0,85	0,95	0,83	0,95
	12	0,74	0,77	0,90	0,90	0,66	0,81	0,85	0,94	0,86	0,95
	13	0,76	0,80	0,89	0,90	0,69	0,81	0,86	0,94	0,86	0,95
	14	0,78	0,77	0,88	0,90	0,70	0,81	0,86	0,94	0,86	0.95

Tabel 5.2 Betrouwbaarheden deeltaken, domeinscores en leerjaar-IQ

		Deeltaken						*Domeinen*			*Leerjaar-*
		FC	*FM*	*GR*	*GA*	*WC*	*WA*	*F*	*G*	*W*	*IQ*
Leerjaar	1	0,72	0,78	0,89	0,91	0,67	0,80	0,84	0,94	0,85	0,95
	2	0,77	0,79	0,88	0,89	0,72	0,81	0,87	0,93	0,86	0,95
	3	0,77	0,81	0,87	0,91	0,70	0,85	0,87	0,94	0,88	0,95

De betrouwbaarheden liggen voor de domeinen waarop gerapporteerd wordt, namelijk Figuren, Getallen en Woorden, voor elke normgroep ruim boven de vereiste 0,80. Ook ligt de betrouwbaarheid van het IQ voor elke normgroep met 0,95 ruim boven de vereiste 0,90. In de tabel zijn voor de volledigheid ook betrouwbaarheden per deeltaak opgenomen, maar deze zijn minder relevant omdat niet op het niveau van afzonderlijke deeltaken wordt gerapporteerd in het leerlingrapport.

5.2 Test-hertestbetrouwbaarheid

Aan de scholen van het normeringsonderzoek is gevraagd of zij in de periode oktober – november van het kalenderjaar waarin zij deelnamen aan het normeringsonderzoek bereid waren deel te nemen aan een test-hertestbetrouwbaarheidsonderzoek. De deelnamebereidheid van de scholen was laag. Uiteindelijk hebben vier scholen deelgenomen aan het testhertestonderzoek met in totaal 221 leerlingen. Voor 200 van deze leerlingen was bekend hoe oud ze waren op het moment van eerste afname.

De verdeling naar onderwijsniveau van deze groep leerlingen (zie tabel 5.3) is niet geheel representatief voor de populatie maar benadert deze voldoende voor dit doel, alleen het speciaal onderwijs ontbreekt in dit test-hertestonderzoek.

Tabel 5.3 Onderzoek test-hertestbetrouwbaarheid: aantallen leerlingen naar onderwijsniveau

Onderwijsniveau	*N*	*%*
VO Brugjaar vmbo	2	1
VO Brugjaar vmbo/havo	4	2
VO Brugjaar havo/vwo	29	13
VO vmbo bb	29	13
VO vmbo kb	58	26
VO vmbo gt	40	18
VO havo	37	17
VO vwo	19	9
Onbekend	3	1
	221	100

Hoewel de aantallen leerlingen bescheiden zijn, hebben we test-hertestbetrouwbaarheden bepaald voor de normgroepen afzonderlijk. Daarbij hebben we als uitgangspunt genomen dat het eerste afnamemoment bepalend was voor de vraag voor welke normgroep de twee resultaten van een leerling meetelden. Een 12-jarige leerling die bij eerste afname van de test in leerjaar 1 zat bijvoorbeeld, telt mee voor de normgroep van de 12 jarigen en voor de normgroep leerjaar 1, ook al zat deze leerling op het moment van hertesten (in oktober) in leerjaar 2. We beschikten over te weinig gegeven om op deze manier de test-hertestgegevens voor leerjaar 3 vast te stellen.
Ofschoon dit test-hertestonderzoek om praktische redenen beperkt van karakter is, geeft het een goede indruk van de test-hertestbetrouwbaarheden, zoals tabel 5.4 laat zien. De test-hertest correlaties voor de

totaalscores liggen voor de onderscheiden normgroepen naar leeftijd en leerjaar tussen de 0,87 en 0,90. Bij de eerder gerapporteerde waarden van 0,95 komen deze waarden in de buurt van wat men zou mogen verwachten of zelfs hoger[9]. Daarnaast moet men het relatief grote interval van ruim een half jaar tussen beide metingen verdisconteren. Normaliter is het interval veel korter, een maand tot zes weken, maar in verband met een te verwachten leereffect is ervoor gekozen dit interval aanzienlijk langer te nemen. Ook voor de domeinscores (F,G en W) zijn de waarden voor de test-hertestbetrouwbaarheden wat lager dan de eerder bepaalde waarden. Gemiddeld liggen de waarden met 0,79 ongeveer 0,10 lager dan de gemiddelde waarde in de overeenkomende groepen in tabel 5.1 en 5.2 (0,88). Alle waarden liggen tussen 0,70 en 0,90 met 0,69 voor figuren in de normgroep 14-jarigen als enige uitzondering.
Voor de volledigheid geven we in tabel 5.5 nog de gemiddelde scores voor de beide testafnames in de onderscheiden normgroepen. De gemiddelde vaardigheidswinst die de kinderen in ongeveer een half jaar tijd laten zien bedraagt zes à acht punten. Dat is aanzienlijk meer dan men op basis van normale ontwikkeling zou verwachten (vergelijk de gemiddelde ruwe scores in tabel 4.10 en 4.11; deze zijn indicatief voor de gemiddelde vaardigheidstoename voor een periode van een jaar). De verschillen wijzen erop dat er sprake is van een duidelijk leereffect bovenop de normale ontwikkeling dat toe te schrijven is aan de herhaalde afname. Het is dan ook niet aan te bevelen om de test voor een tweede maal af te nemen, zeker niet binnen het tijdbestek van een half jaar.

Tabel 5.4 Test-hertestbetrouwbaarheden voor drie leeftijdsgroepen en twee leerjaargroepen

Leeftijd	*Test-hertestcorrelaties*					*Leerjaar*	*Test-herttestcorrelaties*				
	Totaal	*F*	*G*	*W*	*N*		*Totaal*	*F*	*G*	*W*	*N*
12	0,87	0,78	0,71	0,80	39	1	0,88	0,79	0,76	0,83	74
13	0,90	0,82	0,82	0,86	97	2	0,88	0,71	0,81	0,82	147
14	0,88	0,69	0,78	0,81	64	3	--	--	--	--	--
					200						221

Tabel 5.5 Test-hertestonderzoek: gemiddelden in de onderscheiden normgroepen op T1 en T2

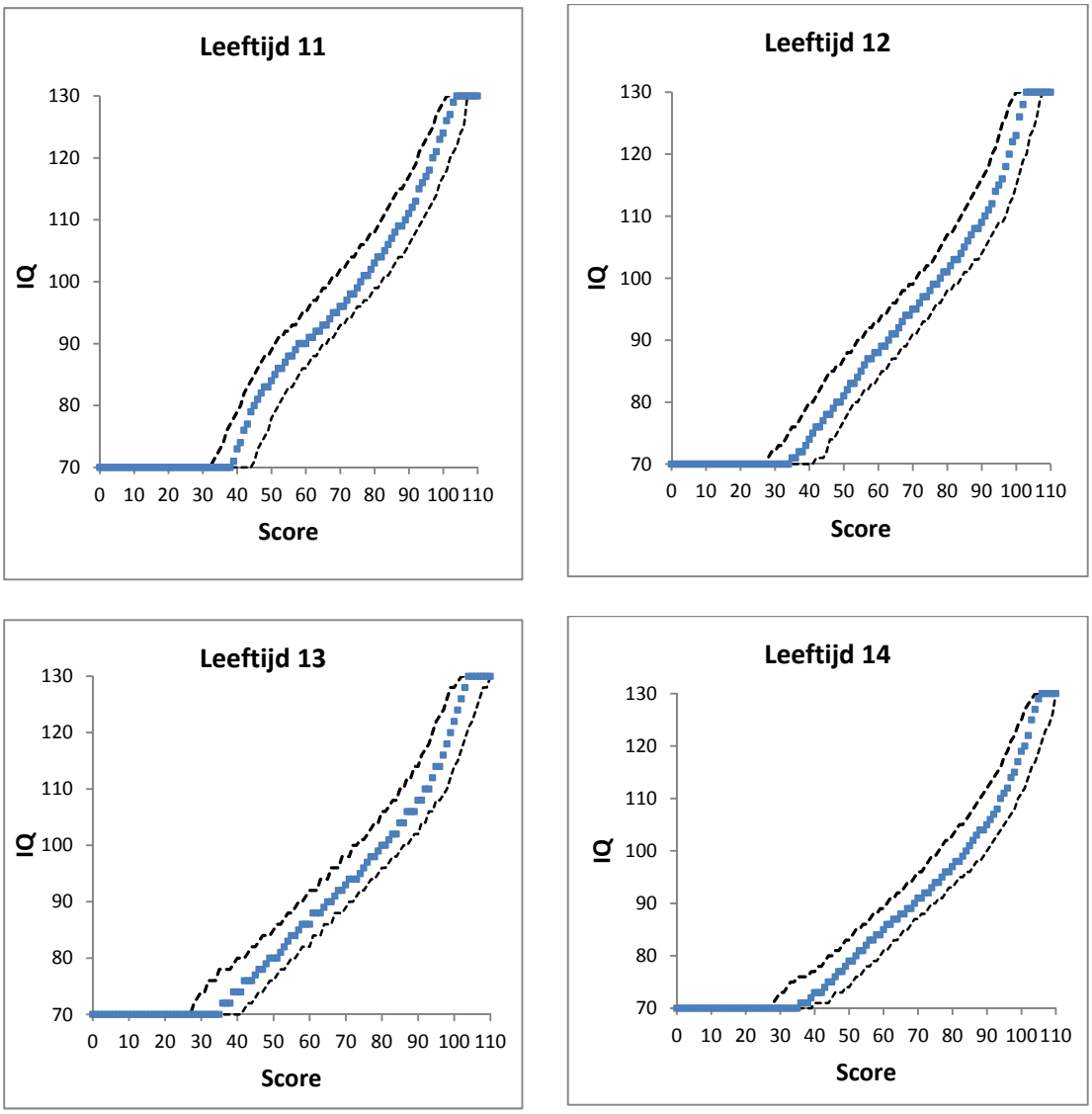
Leeftijd	*Gemiddelden op T1 en T2*					*Leerjaar*	*Gemiddelden op T1 en T2*				
	Totaal	*F*	*G*	*W*	*N*		*Totaal*	*F*	*G*	*W*	*N*
12	77,6 85,5	26,4 29,8	20,2 22,1	31,0 33,6	39	1	74,7 81,7	25,9 28,3	19,2 21,2	29,6 32,2	74
13	80,3 86,2	28,4 30,1	20,6 22,8	31,3 33,3	97	2	82,9 89,2	30,0 32,3	20,8 23,2	32,1 33,7	147
14	81,9 88,7	30,1 32,8	20,5 22,9	31,3 33,0	64	3	--	--	--	--	--
					200						221

5.3 Lokale meetnauwkeurigheid

Figuur 5.4 laat plots zien van de IQ-score op basis van de testscore met 80%- betrouwbaarheidsintervallen rond het leeftijd-IQ en dit voor alle leeftijdgroepen afzonderlijk. De betrouwbaarheidsintervallen zijn bepaald volgens de eerder beschreven Bayesiaanse schattingsprocedure. De plots van figuur 5.5 waarin hetzelfde gebeurt voor het leerjaar-IQ zijn vergelijkbaar. De breedte van het betrouwbaarheidsinterval in termen van het aantal IQ-punten boven en onder een bepaalde score, is af te lezen op de y-as.

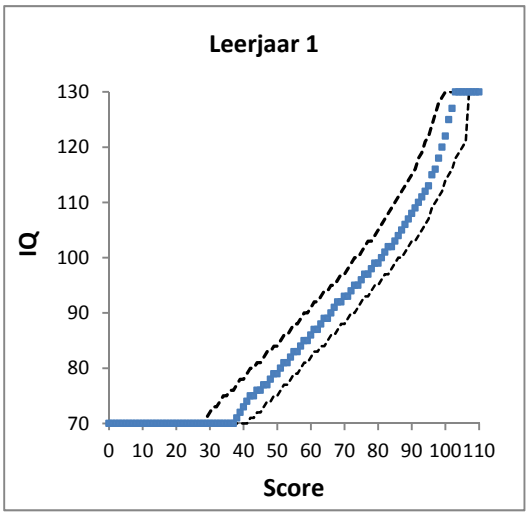
[9] Sijtsma (2012) suggereert op basis van ervaringen in de COTAN-beoordelingspraktijk een verschil van ongeveer 0,10 tussen schattingen van de interne consistentie op basis van alpha en test-hertestbetrouwbaarheden in het nadeel van laatstgenoemde.

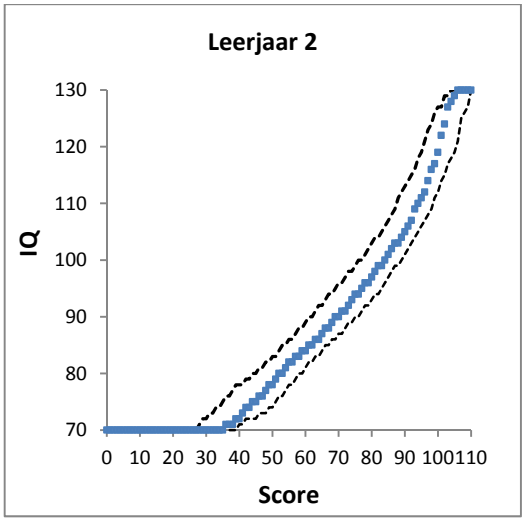
Figuur 5.4 Leeftijd-IQ afgezet tegen de testscore met 80%-betrouwbaarheidsinterval voor vier leeftijdgroepen

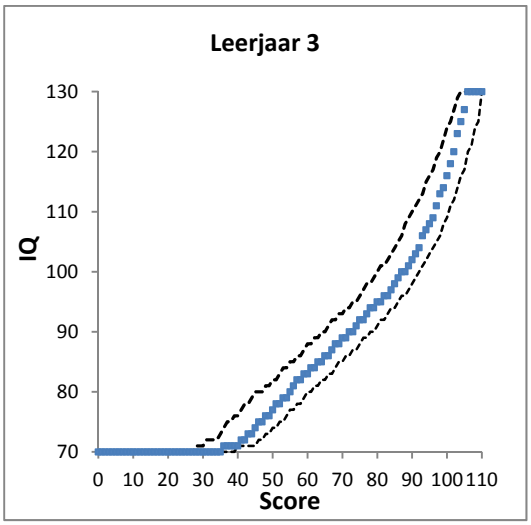


De betrouwbaarheden van het algemene leeftijd- en leerjaar-IQ zijn voor alle normgroepen met 0,95 hoog te noemen (zie de tabellen 5.1 en 5.2). De betrouwbaarheidsintervallen zijn dienovereenkomstig klein. Daarbij moeten we wel de kanttekening plaatsen dat de verschillen tussen kinderen van verschillende leeftijden of leerjaren relatief gering zijn. Het IQ voor de 14-jarigen ligt bijvoorbeeld vrijwel altijd in het betrouwbaarheidsinterval van een 12-jarige met dezelfde score (en *vice versa*).
Verder is het zo dat de betrouwbaarheidsintervallen breder zijn naar mate de score hoger is. Dat is vooral het geval bij de hogere leeftijdgroepen. Door de gekozen (gemiddelde) moeilijkheidsgraad van de items, meet de test wat nauwkeuriger in de lagere en gemiddelde regionen. Overigens blijft dit feit bij veel tests onderbelicht omdat men geen informatie over de lokale meetfout verstrekt. Merk op dat voor de laagste en hoogste scores de betrouwbaarheidsintervallen half open zijn. Dit is een gevolg van het feit dat de laagste en hoogste IQ's niet gedifferentieerd worden gerapporteerd.

Figuur 5.5 Leerjaar-IQ afgezet tegen de testscore met 80%-betrouwbaarheidsinterval voor drie leerjaren







Conclusies met betrekking tot de nauwkeurigheid van het leerjaar-IQ zijn analoog.

6 Validiteit

In dit hoofdstuk rapporteren we onderzoek en analyses met betrekking tot de validiteit van de Intelligentie-test VO. In paragraaf 6.1 komt de begripsvaliditeit aan de orde, in paragraaf 6.2 de criteriumvaliditeit.

6.1 Begripsvaliditeit

De uitgevoerde analyses en resultaten met betrekking tot de begripsvaliditeit zijn te rubriceren volgens een indeling die wordt aangereikt in het COTAN-beoordelingssysteem (Evers et al., 2010). We houden een paragraafindeling aan die hiermee overeenkomt. Dit impliceert dat we achtereenvolgens (1) de dimensionaliteit en structuur van het instrument bespreken, (2) de psychometrische kwaliteit van de testitems, (3) de invariantie van de structuur en itembias, (4) convergente en discriminante validiteit en (5) samenhangen met achtergrondvariabelen en verschillen tussen relevante subgroepen.

Dimensionaliteit en structuur
In hoofdstuk 4 hebben we uiteengezet hoe de test is geconstrueerd en welk IRT-model we daarbij hebben gehanteerd, te weten het marginale multidimensionele One-Parameter Logistic Model (OPLM). Kalibratie is geschied per deeltaak (FC, FM, GR, GA, WC en WA) en per leeftijdsgroep, waarbij we aan de hand van R1c-toetsen konden aantonen dat de modelpassing goed is. Op grond van de itemselectie in de constructiefase en de daarbij aangetroffen hoge gemiddelde R_{it}-waarden (zie hoofdstuk 3) was dit naar verwachting. Op grond van deze analyses kunnen we dus vaststellen dat we bij de meting van de algemene intelligentie gebruikmaken van deeltaken die ieder een unidimensionele (latente) vaardigheid representeren.

Overeenkomstig onze visie op algemene intelligentie beogen we een sommering over deeltaakscores. Om deze sommering op zinvolle wijze te kunnen uitvoeren is het nodig dat de deeltaakscores onderling voldoende samenhangen en dat er sterke samenhangen zijn tussen de deeltaakscores en de somscore (testscore). Eerder hebben we al vermeld dat de intercorrelaties voor de deeltaakscores hoog tot zeer hoog zijn. We hebben deze correlaties per leeftijdsgroep berekend omdat er per leeftijdgroep is gekalibreerd; zij variëren tussen 0,56 en 0,90. In aanvulling hierop geven we in tabel 6.1 eveneens per leeftijdscategorie de correlaties tussen de deeltaakscores en de testscore. De correlaties zijn hoog en liggen voor vijf van de zes deeltaken in de orde van grootte van 0,79 tot 0,85. Alleen voor de deeltaak WC is de samenhang wat lager (0,67 – 0,75), maar toch altijd nog aanzienlijk. De gevonden samenhangen beantwoorden aan onze verwachtingen en rechtvaardigen het sommeren over deeltaken. Bovendien kunnen we vaststellen dat de correlaties voor de onderscheiden leeftijdsgroepen een vrijwel gelijk patroon laten zien.

Tabel 6.1 Correlaties tussen deeltaakscores en de testscore

		Deeltaak					
		FC	FM	GR	GA	WC	WA
Leeftijd	11	0.82	0.81	0.83	0.86	0.70	0.78
	12	0,79	0,80	0,85	0,85	0,67	0,82
	13	0,79	0,80	0,83	0,84	0,72	0,79
	14	0,79	0,81	0,84	0,84	0,75	0,82

Ten slotte valt over de structuur nog op te merken dat we sommeren over drie domeinen, namelijk Figuren, Getallen en Woorden. Zo'n sommering per domein impliceert dat de correlatie tussen de twee deeltaakscores per domein steeds hoger is dan de correlaties tussen deeltaken die niet tot hetzelfde domein behoren. Ook hierop zijn we in hoofdstuk 4 al ingegaan, waarbij we aan de hand van figuur 4.9 op

grafische wijze hebben laten zien dat dit patroon zich inderdaad voordoet. Daarbij hebben we aangegeven dat dit bij alle leeftijdgroepen het geval is. In tabel 6.2 vatten we uitkomsten van de analyses nog eens samen, waarbij we ons vanwege de overzichtelijkheid beperken tot de mediane waarden over de leeftijdgroepen (het patroon verschilt immers niet over de leeftijdgroepen). De correlaties tussen deeltaken binnen hetzelfde domein liggen alle boven de 0,80, met een gemiddelde van 0,83. De correlaties tussen deeltaken van verschillende domeinen liggen alle onder de 0,80 met een gemiddelde van 0,71

Tabel 6.2 Correlaties tussen deeltaken (mediane waarden over leeftijdgroepen)

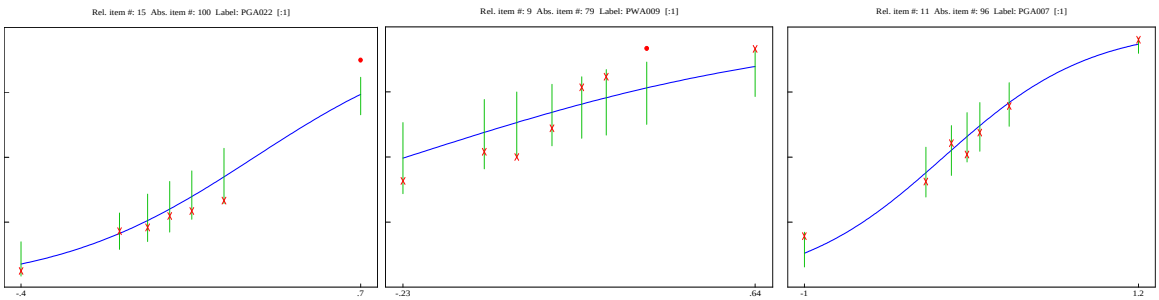
	Deeltaak				
	FM	*GR*	*GA*	*WC*	*WA*
FC	**0,83**	0,75	0,73	0,76	0,72
FM		0,72	0,77	0,67	0,71
GR			**0,81**	0,65	0,68
GA				0,64	0,69
WC					**0,86**

Psychometrische kwaliteit van de items

Ook over de psychometrische kwaliteit van de testitems is in het voorafgaande al het nodige gezegd. Immers, de geslaagde kalibratie op basis van het IRT-model zorgt ervoor dat de items per deeltaak steeds een unidimensionele latente vaardigheid representeren, wat hoge R_{ir}-waarden impliceert. We vermeldden hierover in hoofdstuk 4 al, dat bij de R1c-toetsingen per deeltaak waar de passing iets minder bleek te zijn, een zeer gering aantal items voor dit relatieve gebrek aan passing verantwoordelijk was.

We gaan hier nog wat nader in op de twee items die een minder goede passing lieten zien. Het betreft hier een item van de deeltaak GA dat minder goed paste voor de leeftijdgroep van de 14-jarigen (meest linkse afbeelding in figuur 6.1) en een item van deeltaak WA dat minder goed paste voor de leeftijdgroep van de 11-jarigen (middelste afbeelding). We geven informatie voor beide items in de vorm van empirische itemresponse-curves (vergelijk de procedure die we hebben beschreven in hoofdstuk 4 en de voorbeelden in figuur 4.1 en 4.2). Ter vergelijking voegen we soortgelijke gegevens toe voor een goed passend item (GA voor 12-jarigen, meest rechtse afbeelding).

Figuur 6.1 Empirische itemresponse-curves voor twee minder goed passende items met rechts een goed passend item ter vergelijking (toelichting in de tekst)



Het is goed te zien dat de eerste twee items een minder goede *fit* laten zien. Bovendien discrimineert het WA-item relatief slecht. Toch valt het gebrek aan passing mee. Het gaat immers steeds om slechts één van

de vier kalibraties die per deeltaak werden uitgevoerd (i.e. voor elke leeftijdsgroep afzonderlijk) en bovendien vallen de waarden (rood aangegeven in de figuur) op één na binnen de aangegeven intervallen (aangeduid met de groene balkjes). Inspectie van de uiteindelijke R_{ir}-waarden liet zien dat alle items voldoende tot goed discrimineerden.
We kunnen concluderen dat de itemkwaliteit hoog is: de items discrimineren goed en laten een goede modelpassing zien. Dat was op basis van het constructie-onderzoek uiteraard al duidelijk. We vatten hier de uitkomsten van dit in hoofdstuk 3 gerapporteerde onderzoek nog eens samen.
De gemiddelde R_{it}-waarden voor de items bleken zeer hoog zijn en als 'goed' te kwalificeren; ze variëren tussen 0,34 en 0,55. Slechts twee testitems (1,8%) voldoen niet aan de grenswaarde voor 'voldoende' van 0,20 die door de COTAN wordt aangegeven, terwijl 90% van de items de kwalificatie 'goed' verdient.
Ook voor de p–waarden geldt dat deze voldoen aan de uitgangspunten die bij de testconstructie daarvoor werden vastgelegd (zie Tabel 3.2). Gemiddeld liggen de p-waarden voor vijf van de zes deeltaken net iets boven de 0,70, met een uitzonderingspositie voor WC (gemiddelde p-waarde 0,82).

Invariantie van de structuur en itembias (DIF)
Eerder hebben we aangegeven dat de testitems bij kalibratie DIF naar leeftijd lieten zien. De conclusie was dat leerlingen van verschillende leeftijd en verschillend ontwikkelingsniveau door de testitems voor kwalitatief verschillende problemen worden geplaatst. Daarom zijn de kalibraties uiteindelijk per leeftijds-groep uitgevoerd.

Voor intercorrelaties tussen de deeltaakscores enerzijds en voor de correlaties tussen deeltaakscores en de testscore anderzijds geldt echter dat het patroon (de structuur) voor de verschillende leeftijdsgroepen in grote lijnen gelijk is. Hetzelfde geldt voor groepen die te vormen zijn op basis van leerjaar. We verwijzen hier naar de eerder gepresenteerde gegevens die laten zien dat de structuur van de intercorrelaties overeenkomt met de verwachtingen en een hoge mate van invariantie laat zien voor leeftijd- en leerjaar-groepen.

Convergente en divergente validiteit
We hebben ons veel moeite getroost om de soortgenootvaliditeit van de Intelligentietest VO in kaart te brengen. In eerste instantie deden we dat door aan de scholen die aan het normeringsonderzoek deelnamen te vragen om ook leerlingscores op andere instrumenten door te geven, te weten (voor zover bekend) de score op de Eindtoets en / of op de NIO (Nederlandse Intelligentietest voor Onderwijsniveau). Slechts een beperkt aantal scholen was bereid die informatie te verstrekken. Belangrijkste reden die door scholen werd opgeven was gebrek aan tijd en druk zijn met andere werkzaamheden. In totaal gaven slechts acht scholen leerlingscores door van de NIO (vijf scholen) en / of Eindtoets (zeven scholen).
De onderwijsniveaus van de scholen voor voortgezet onderwijs varieerden van vmbo bb tot en met vwo. Ook enkele scholen voor basisonderwijs leverden gegevens aan.

In de periode mei-juni 2009 is aanvullend onderzoek gedaan naar de soortgenootvaliditeit. Er werden scholen benaderd die naast de Eindtoets in 2009 ook de in 2009 uitgegeven Intelligentietest Eindtoets Basisonderwijs (IT-EB) van het Cito hadden gemaakt. In totaal deden 33 scholen mee met aan dit onderzoek. Bij 736 leerlingen van deze scholen werd de Intelligentietest VO afgenomen. Van deze leerlingen waren er 520 die ook de Cito IT-EB hadden gemaakt. Van 706 leerlingen kon een leeftijd-IQ worden vastgesteld en van 609 van deze leerlingen is ook de Eindtoetsscore bekend. Voor alle instrumenten die gebruikt zijn in het kader van dit onderzoek naar de soortgenootvaliditeit geldt dat de psychometrische kwaliteit uitstekend is. De COTAN-beoordeling voor de NIO is op alle criteria goed. Hetzelfde geldt voor de Eindtoets Basisonderwijs met uitzondering van de begripsvaliditeit die als voldoende is beoordeeld. Voor de IT-EB zijn de uitgangspunten bij de testconstructie en de begrips- en criteriumvaliditeit als voldoende beoordeeld, de overige criteria als goed.

De via het normeringsonderzoek en het aanvullende onderzoek verzamelde gegevens werden ten behoeve van de analyses samengevoegd. De resultaten van de analyses en de bijbehorende leerlingaantallen zijn weergegeven in tabel 6.3.

Tabel 6.3 Soortgenootvaliditeit: samenhangen tussen diverse instrumenten die gebruikt worden ten behoeve van plaatsing van leerlingen in verschillende typen vervolgonderwijs in het VO

	Leeftijd-IQ	*Eindtoets*	*IT-EB*	*NIO*
Leerjaar-IQ	0,99	0,74	--	0,89
	N=2186	N=175		N=55
Leeftijd-IQ		0,76	0,74	0,81
		N=761	N=511	N=105
Eindtoets (EB)			0,72	--
			N=520	

De in de tabel vermelde correlatie tussen leeftijd-IQ en leerjaar-IQ is gebaseerd op de leerlingen die aan het normeringsonderzoek deelnamen. Omdat het leerjaar-IQ alleen voor het voortgezet onderwijs kon worden berekend, betreft het hier uitsluitend vo-leerlingen. Zoals te verwachten is de correlatie van 0,99 tussen de beide waarden zeer hoog en maakt het in de regel weinig of niets uit welk IQ men in de praktijk hanteert. Voor verreweg de meeste leerlingen zal het leerjaar- en leeftijd-IQ exact gelijk zijn.
Het meest interessant vanuit het oogpunt van soortgenootvaliditeit zijn de correlaties met de NIO, vooral die met het leerjaar-IQ (omdat de NIO naar leerjaar is genormeerd). De correlaties zijn met 0,89 en 0,81 hoog en bevredigend te noemen. De correlatie van 0,81 heeft betrekking op 105 leerlingen in het basisonderwijs voor wie alleen een leeftijd-IQ kon worden berekend. Dat de correlatie voor basisschoolleerlingen enigszins lager uitvalt, kan ermee te maken hebben dat de Intelligentietest VO qua aard en niveau meer is afgestemd op leerlingen in het voortgezet onderwijs. Ook de correlatie van 0,74 tussen de beide Cito intelligentietesten (beide afgenomen bij basisschoolleerlingen) wijst hierop. Daarnaast kent de Intelligentietest IB ook een wat andere samenstelling. De taken van deze test doen minder uitsluitend een beroep op redeneervaardigheden en doen relatief wat meer een beroep op kennis die op school wordt verworven (met name bij de subtests Synoniemen en Tegenstellingen).
De correlaties tussen leerjaar-IQ (175 leerlingen van het VO) en leeftijd-IQ (761 leerlingen van het basisonderwijs) enerzijds en de score op de Eindtoets Basisonderwijs anderzijds beantwoorden met 0,74 en 0,76 aan de verwachtingen. Beide instrumenten, Intelligentietest VO en Eindtoets worden gebruikt om de advisering omtrent de plaatsing in het voortgezet onderwijs te ondersteunen, maar doen een beroep op duidelijk verschillende vaardigheden. In de Eindtoets gaat het bij uitstek om gerealiseerde leervorderingen, in de Intelligentietest VO uitsluitend om redeneervaardigheden die relatief weinig zijn beïnvloed door wat er op school aan onderwijs is aangeboden.
Bij laatstgenoemde samenhangen kunnen we aantekenen dat deze ook te duiden zijn in termen van criteriumvaliditeit. In paragraaf 6.2 komen we dan ook op deze cijfers terug.

Ook de samenhangen met scores op de leervorderingentoetsen van het Cito Volgsysteem voortgezet onderwijs zijn op te vatten als relevant voor de criteriumvaliditeit. We beschouwen ze in deze paragraaf vooral vanuit het perspectief van de convergente en discriminante validiteit. Drie scholen voor voortgezet onderwijs uit het normeringsonderzoek hebben we bereid gevonden om gegevens van het Cito Volgsysteem voortgezet onderwijs aan te leveren. Dit volg-en adviessysteem bevat een aantal leervorderingentoetsen voor enkele voor de eerste drie leerjaren van het voortgezet onderwijs belangrijke vaardigheden. Het betreft toetsen voor Nederlandse leesvaardigheid, Nederlandse woordenschat, rekenen-wiskunde, studievaardigheden en Engels. De psychometrische gegevens voor deze toetsen zijn prima. Zij zijn op alle criteria door de COTAN als ‘goed’ beoordeeld, met uitzondering van de begripsvaliditeit (‘voldoende’); de op grond van ontbrekend onderzoek als ‘onvoldoende’ beoordeelde criteriumvaliditeit is hier niet zo relevant.
De gegevens die we rapporteren in tabel 6.4 zijn afkomstig van leerlingen uit leerjaar 1. Alleen de onderwijsniveaus havo/vwo en vwo zijn vertegenwoordigd. Dit betekent dat er sprake is van een aanzienlijke ‘restriction of range’ in vergelijking met een situatie waarin we ook gegevens van scholen uit het vmbo hadden kunnen meenemen. Dit impliceert over het algemeen dat de correlaties lager uitvallen.
Van 395 leerlingen zijn een of meer scores op de genoemde leervorderingentoetsen beschikbaar. Voor de toetsen Nederlands en Rekenen-wiskunde zijn de correlaties met de IQ-indicatoren gebaseerd op ongeveer

350 leerlingen. Voor de toetsen Engels en Studievaardigheden zijn de correlaties met de IQ-indicatoren gebaseerd op ongeveer 180 leerlingen.
In de tabel zijn vooral de drie rijen met samenhangen tussen de percentielscores voor Figuren, Woorden en Getallen enerzijds en de toetsscores anderzijds van belang. Daarnaast zijn ook de samenhangen tussen de algemene intelligentiescores (leeftijd- en leerjaar-IQ) en de toetsen opgenomen. Voor de volledigheid staan ook de correlaties tussen de toetsscores onderling in de tabel.

Tabel 6.4 Convergente en discriminante validiteit: samenhang) *tussen intelligentie-indicatoren en leervorderingentoetsen in het Cito Leerlingvolg- en adviessysteem VO*

	NL leesvaardigheid	*NL woordenschat*	*Rekenen-wiskunde*	*Studie-vaardigheden*	*Engels*
Leeftijd-IQ	0,37	0,36	0,62	0,45	0,28
Leerjaar-IQ	0,39	0,37	0,64	0,49	0,33
Percentielscore Figuren	0,25	0,26	0,48	0,29	0,17
Percentielscore Woorden	0,43	0,38	0,38	0,19	0,27
Percentielscore Getallen	0,27	0,17	0,46	0,25	0,18
NL leesvaardigheid	x	0,43	0,34	0,33	0,44
NL woordenschat		x	0,36	0,28	0,43
Rekenen-wiskunde			x	0,49	0,28
Studievaardigheden				x	0,29
Engels					x

*) gebaseerd op een homogene groep havo-vwo-leerlingen

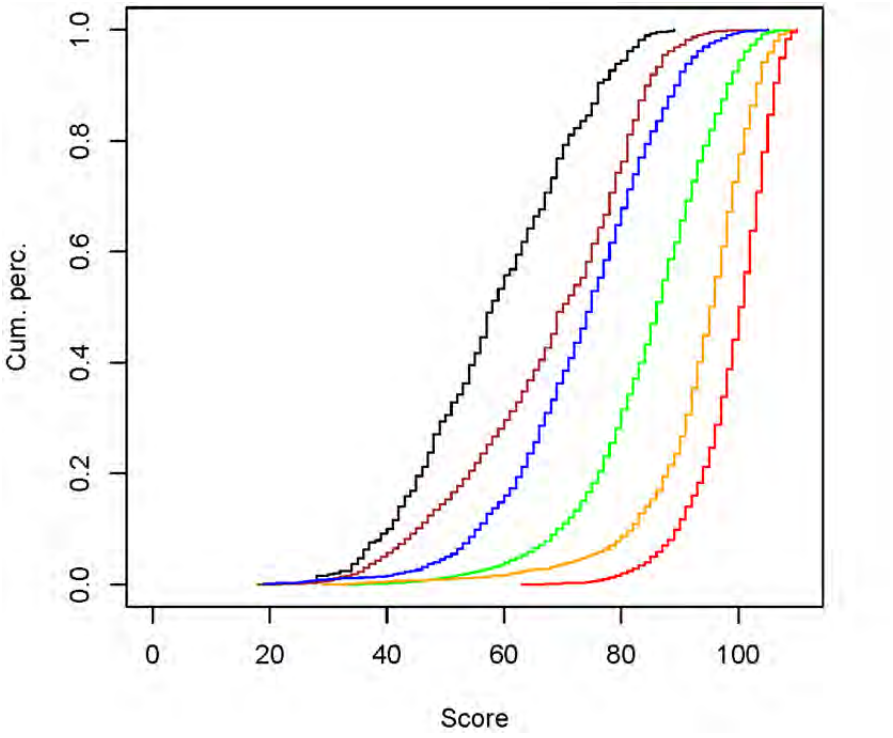
Correlaties waarvoor het mogelijk was om op voorhand uitspraken te doen over de te verwachten hoogte (in vergelijki ng met andere correlaties) zijn in de tabel gearceerd. Wanneer we deze langslopen hadden we de verwachting dat de scores op de Intelligentietest VO het sterkst zouden samenhangen met de toetsscore op Rekenen-wiskunde. Deze verwachting is gebaseerd op het zware accent dat in de test wordt gelegd op redeneervaardigheden, dit laatste ongeacht het domein van redeneren. We zien dat onze veronderstellingen empirische steun ondervinden met correlaties van boven 0,60 die er in vergelijking met andere toetsscores duidelijk uitspringen. De overige verwachtingen hadden betrekking op verschillen in samenhang tussen toetsscores voor talige onderdelen en rekenen-wiskunde enerzijds en de drie domeinscores (Figuren, Woorden en Getallen) anderzijds. Voor studievaardigheden hadden we vooraf geen duidelijke hypothese. Voor de drie talige onderdelen verwachtten we de hoogste correlatie met het domein Woorden. Voor Rekenen-wiskunde daarentegen verwachtten we de hoogste correlatie met Figuren en Getallen. Alle verwachtingen worden bevestigd door de data, al zijn de verschillen tussen de correlaties niet al te groot.

Samenhang met achtergrondvariabelen, verschillen tussen groepen
In hoofdstuk 4 zijn we al ingegaan op verschillen in gemiddelden die optreden bij de onderscheiden normgroepen. We zagen te verwachten verschillen naar leeftijd en naar leerjaar (reden waarom juist voor deze groepen afzonderlijke normen zijn opgesteld). Daarnaast zijn ook andere subgroepen te vormen waarvoor het interessant is om na te gaan in hoeverre hun gemiddelden van elkaar verschillen. In het onderstaande gaan we achtereenvolgens in op verschillen naar onderwijstype en –niveau, verschillen naar thuistaal en verschillen naar sekse.

Verschillen naar onderwijstype en -niveau
Eén van de belangrijkste doelen van de Intelligentietest VO is het bieden van ondersteuning bij de advisering omtrent de meest geschikte plaatsing van leerlingen in het voortgezet onderwijs. Daarbij is het uitgangspunt bekend: de in het voortgezet onderwijs onderscheiden typen en niveau's doen een beroep op

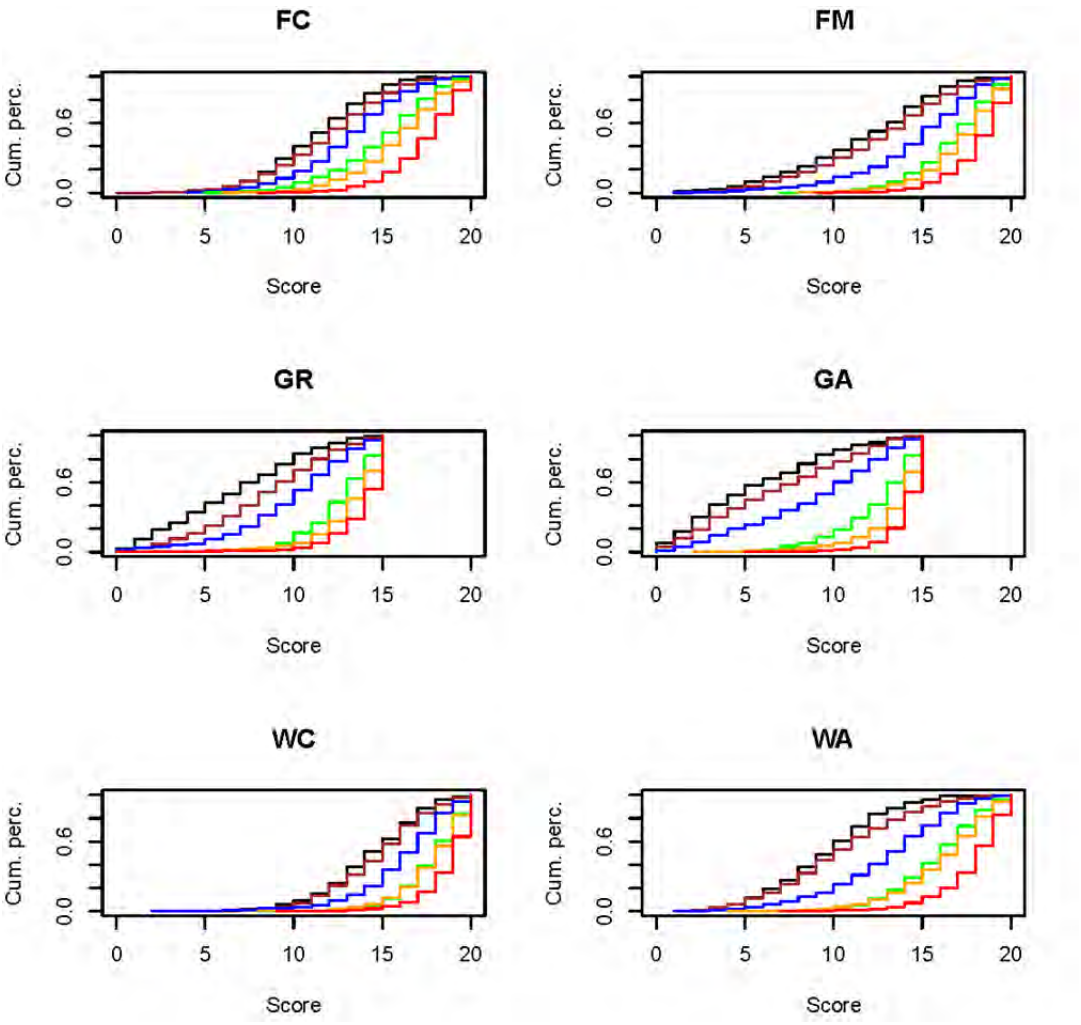
kwantitatief verschillende niveaus van cognitieve vaardigheden. Daarbij nemen we aan dat deze verschillen in cognitieve vaardigheden tot uitdrukking komen in het functioneren op de intelligentietest.

Figuur 6.2 Verdeling van de testscore naar schooltype in leerjaar 3: bb+, bb, kb, gt, havo, vwo



Als we verschillen naar schooltype- en –niveau duidelijk in beeld willen brengen, kunnen we ons het beste beperken tot leerjaar 3. Op dat moment hebben immers vrijwel alle leerlingen een plaats gekregen in een bepaald school- en onderwijstype, zoals bb+, basis- en kaderberoepsgerichte en gemengd theoretische leerweg in het vmbo, havo en vwo (in het vervolg af te korten tot bb+, bb, kb, gt, havo en vwo)
Figuur 6.2 toont de scoreverdelingen voor elk type. De verdelingen liggen geordend zoals verwacht: de bb+-leerlingen (zwarte curve) scoren het laagst en de vwo-leerlingen (rode curve) het hoogst. De overige curven liggen hiertussen in de verwachte volgorde: bb, kb, gt en havo). De vwo-leerlingen scoren in meerderheid boven de 90 punten (dat wil zeggen, zij behalen minimaal 80 procent van de maximumscore). Met name voor de bb+-, bb- en kb-kinderen is de test niet gemakkelijk.

Figuur 6.3 Verdelingen van deeltaakscores naar type: bb+, bb, kb, gt, havo, vwo



Een uitsplitsing naar deeltaken is te zien in figuur 6.3. Ook hier zijn de scoreverdelingen geordend zoals verwacht. Wellicht interessant is dat de vwo-leerlingen zich het meest onderscheiden op FC, WC en WA. Bij Woorden is er nauwelijks verschil tussen bb+ en bb onderling, en tussen havo en gt onderling; havo- en gt-leerlingen onderscheiden zich vooral op Getallen. Het is lastig om aan deze en andere verschillen gerichte interpretaties te verbinden. Daarom beperken we ons tot een simpele conclusie in relatie tot de validiteit van de test: de scoreverdelingen op deeltaken naar type gedragen zich zoals bedoeld en verwacht.

Verschillen naar thuistaal

Voor de leerlingen in de verschillende onderzoeken werd nagegaan, welke taal het kind thuis met de ouders/verzorgers sprak. Daarbij kon onderscheid gemaakt worden tussen Nederlands, Nederlandse streektaal/dialect, Fries, andere West-Europese taal, Oost-Europese taal, Turks, Marrokaans-Arabisch, Berbers/Tamazight, Surinaams, Hindoestaans, Papiaments en overig.
De leerlingen die Nederlands, Nederlandse streektaal/dialect of Fries hebben aangegeven zijn samengevoegd tot de categorie Nederlands. De overige leerlingen zijn samengevoegd tot de categorie Niet-Nederlands.

Op voorhand was de verwachting dat leerlingen met een niet-Nederlandse thuistaal lager zouden scoren dan leerlingen die thuis Nederlands spreken. Deze verwachting is niet zo zeer gebaseerd op de taligheid van de test als wel op de achtergrond en situatie van het gezin waar de leerling deel van uit maakt.
De taken hebben immers voor een aanzienlijk deel betrekking op niet-talige redeneervaardigheden

(Figuren, Getallen). Voor zover de inhoud wél talig van aard is (Woorden) is er gekozen voor veelal eenvoudige en hoogfrequent voorkomende woorden. Aan de andere kant is niet goed in te schatten wat de achtergronden zijn van de gezinnen waarin thuis geen Nederlands wordt gesproken. Het is mogelijk dat deze leerlingen nog niet zo lang in Nederland verblijven of nog weinig ingeburgerd zijn, dat de ouders een laag opleidingsniveau hebben en een daarmee samenhangende lage sociaal-economische status, of een combinatie van dit soort kenmerken.

Tabel 6.5 Verschillen tussen kinderen die thuis al dan niet Nederlands spreken

Schooltype / onderwijsniveau	*Thuistaal Nederlands*			*Thuistaal niet-Nederlands*			*T-test*		
	Gem.	*SD*	*N*	*Gem.*	*SD*	*N*	*T*	*df*	*p*
Basisonderwijs									
Leerjaar 6, 7 en 8 regulier	102,6	13,85	889	98,8	11,95	47	1,843	934	*
VO brugklassen									
Brugjaar vmbo	90,2	11,18	186	81,5	9,15	41	4,650	225	***
Brugjaar vmbo/havo	100,4	10,64	161	97,5	15,44	6	0,643	165	ns
Brugjaar havo/vwo	110,6	10,48	415	106,5	10,32	96	3,473	509	***
VO categoriaal									
vmbo-bb	85,6	9,25	241	83,2	8,42	104	2,351	343	**
vmbo-kb	94,2	8,57	240	88,3	8,91	15	2,599	253	***
vmbo-gt	101,3	11,99	260	94,9	13,27	16	2,078	274	*
havo	111,5	10,57	140	102,7	11,38	54	5,102	192	***
vwo	118,8	9,08	528	111,0	9,40	91	7,582	617	***

* p<0,05; ** p<0,01; ***p<0,001

De verschillen in gemiddelde tussen beide groepen werden per onderwijstype eenzijdig getoetst. In alle gevallen scoorden de leerlingen die thuis Nederlands spreken gemiddeld hoger dan de leerlingen die dit niet doen. Alle verschillen zijn significant, met uitzondering van de toetsing voor brugjaar vmbo. Opvallend hoog zijn de verschillen voor de hoogste onderwijsniveaus, havo en vwo.
Het is niet duidelijk wat de exacte achtergrond van de verschillen is. Ook is het niet bekend welke rol de toewijzing aan de verschillende onderwijsniveaus in het voortgezet onderwijs speelt. Wél duidelijk is het dat het hier om een systematisch effect gaat; het is dus niet zo dat de effecten alleen voorkomen in bepaalde niveaus. Hoe dan ook, men doet er verstandig aan het feit dat een leerling thuis geen Nederlands spreekt in de overwegingen te betrekken om de test al dan niet af te nemen en, als de test wordt afgenomen, dit gegeven in de interpretatie van de resultaten te betrekken.

Verschillen naar sekse
In tabel 6.6 zijn de gemiddelde IQ-scores voor jongens en meisjes gegeven, berekend op basis van alle kinderen bij wie de test in het kader van de normerings- en valideringsonderzoeken is afgenomen (N = 4015). Deze dataset valt niet geheel samen met de normeringssteekproef: er is relatief vaak sprake van leerlingen in hogere onderwijsniveaus en bovendien is er geen sprake van weging. Daardoor vallen de gemiddelden hoger uit dan men op basis van de normeringsgemiddelden (=100) zou verwachten.
Het leerjaar-IQ is alleen berekend voor leerlingen in het voortgezet onderwijs (N = 2724). Zowel voor het leeftijd-IQ als voor het leerjaar-IQ zijn de verschillen tussen jongens en meisjes niet significant. Wel valt op dat de standaarddeviatie voor het leeftijd-IQ bij de meisjes wat lager is dan bij de jongens. De effectgroottes voor beide IQ-scores zijn erg klein (respectievelijk 0,04 en 0,07 voor leeftijd- en leerjaar-IQ) en leiden tot de conclusie dat er voor sekse sprake is van de bij de aangegeven effectgroottes passende interpretatie (Cohen, 1992): “geen effect”.

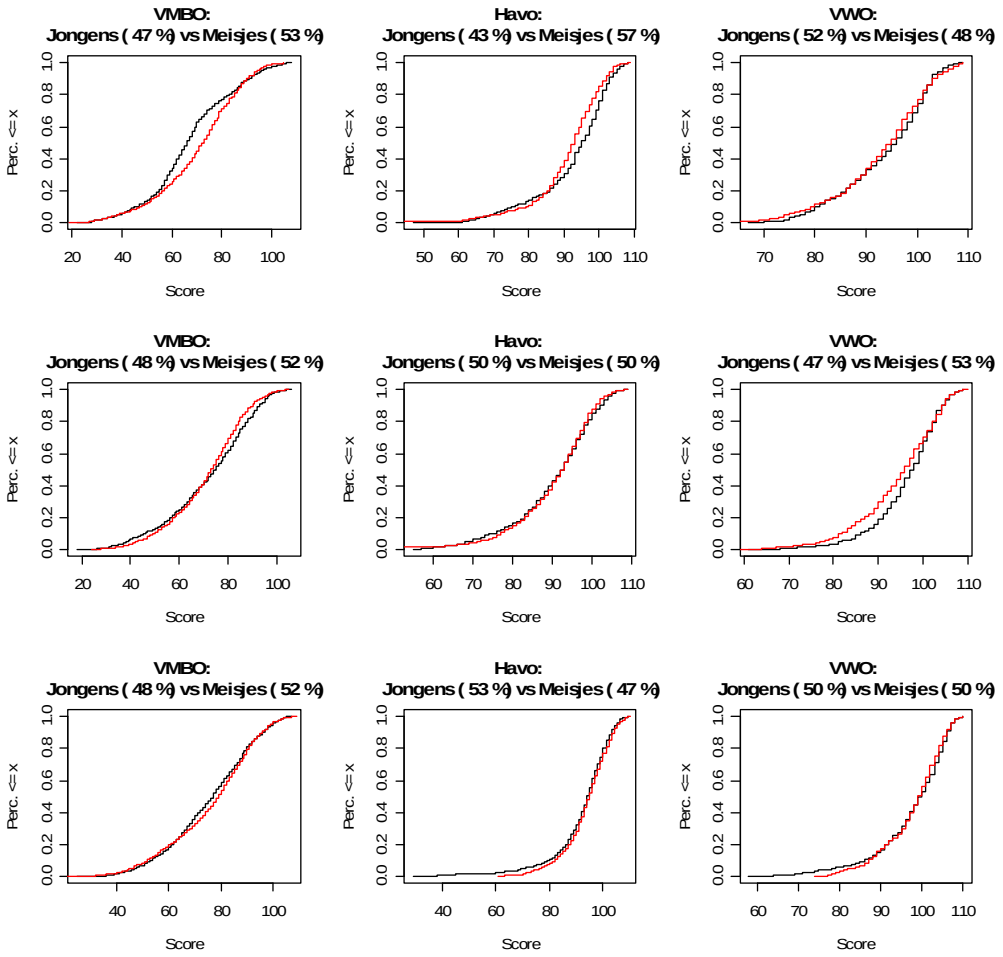
Tabel 6.6 *Verschillen naar sekse voor leeftijd-IQ (N=4015) en leerjaar-IQ (exclusief basisonderwijs, N=2724)*

					T-test	
	sekse	*N*	*gemiddelde*	*SD*	*T*	*df*
leeftijd-IQ	jongens	1898	103,6	15,46	1,225*)	3931
	meisjes	2017	103,0	14,35		

	sekse	N	gemiddelde	SD		
leerjaar-IQ	jongens	1335	103,4	15,71	1,891*)	2722
	meisjes	1389	102,2	15,34		

*) niet significant

Figuur 6.4 *Verdelingen van de testscore voor jongens en meisjes van 12, 13 en 14 jaar (de rijen) in verschillende schooltypen (de kolommen). De curve van de meisjes is rood, die van de jongens zwart.*



In aanvulling op deze analyse zijn in figuur 6.4 de scoreverdelingen getekend van 12-, 13- en 14-jarige meisjes en jongens voor verschillende schooltypen binnen het voortgezet onderwijs (11-jarigen zijn buiten beschouwing gelaten vanwege hun geringe aantal binnen het VO). Deze zijn afgeleid van de normgroepen. We verwachtten geen systematische verschillen tussen jongens en meisjes en we zien dat die ook vrijwel afwezig zijn. Bij de 12-jarigen op het vmbo doen de meisjes het iets beter, terwijl op het vwo 13-jarige jongens het iets beter doen. Op elk schooltype komen meisjes ongeveer even vaak voor als jongens.

Op grond van deze analyses is de conclusie gerechtvaardigd dat de gemiddelde testscores van jongens en meisjes elkaar nauwelijks ontlopen.

Dyslexie
Er is nagegaan in hoeverre het al dan niet dyslectisch zijn van invloed is op de testscore. Voor deze analyses is de volledige, niet gewogen dataset gebruikt. Op voorhand werden geen of slechts geringe verschillen verwacht. De test doet immers vooral een beroep op redeneervaardigheden waarvan men mag aannemen dat deze voor dyslectici niet slechter ontwikkeld zijn dan voor andere leerlingen. Bovendien zijn de taken grotendeels niet-talig van aard, met uitzondering van het domein Woorden. In de taken is echter gekozen voor relatief eenvoudige woorden met een hoge frequentie van voorkomen.
In tabel 6.7 zijn op de eerste plaats de verschillen voor de groep basisschoolleerlingen weergegeven, waarbij we ons moeten beperken tot het leeftijd-IQ. Deze gegevens zijn het eenvoudigst te interpreteren omdat eventuele effecten van verschillen in plaatsing in het voortgezet onderwijs bij deze subgroep geen rol spelen. Dyslectici en niet-dyslectici blijken nauwelijks en niet-significant van elkaar te verschillen; er is geen sprake van een effect (Cohens d = 0,03) .

Tabel 6.7 IQ-scores voor dyslectici en niet-dyslectici

	basisonderwijs				*T-test*	
	dyslexie	*N*	*gemiddelde*	*SD*	*T*	*df*
leeftijd-IQ	niet-dyslectisch	909	102,4	13,81	0,205*)	967
	dyslectisch	60	102,0	14,08		
	*voortgezet onderwijs***)*					
	dyslexie	*N*	*gemiddelde*	*SD*		
leerjaar-IQ	niet-dyslectisch	1941	104,2	15,47	0,600*)	2011
	dyslectisch	72	103,1	16,28		
leeftijd-IQ	niet-dyslectisch	1941	104,8	15,68	0,727*)	2011
	dyslectisch	72	103,4	16,66		

*) niet significant

**) alleen leerlingen voor wie zowel een leerjaar-IQ als een leeftijd-IQ kon worden berekend

Wanneer we ook de VO-leerlingen in de analyses betrekken, kunnen we zowel naar leerjaar- als naar leeftijd-IQ kijken. Daarbij hebben we ons beperkt tot de groep leerlingen voor wie zowel het leerjaar- als het leeftijd-IQ kon worden berekend (dit om de onderlinge vergelijkbaarheid te verhogen). Het voor basisschoolleerlingen – aan de hand van het leeftijd-IQ - gerapporteerde beeld wordt voor leerlingen in het voortgezet onderwijs bevestigd. De verschillen tussen dyslectici en niet-dyslectici zijn ook hier klein en niet significant. Dat geldt ook wanneer we de verschillen tussen beide groepen bekijken aan de hand van het leerjaar-IQ.
Op basis van het bovenstaande kan worden geconcludeerd dat leerlingen met en zonder dyslexie geen significante verschillen laten zien in hun score op de intelligentietest. De test kan dus bij dyslectici zonder bezwaren worden toegepast.

6.2 Criteriumvaliditeit

Bij criteriumvaliditeit gaat het in principe om de vraag of de test een goede voorspeller is van “niet-testgedrag (retrospectief, gelijktijdig of predictief)” (zie Evers et al., 2010). Daarbij zal de aard van het criterium vooral bepaald worden door de doelen en functies van het instrument. Bij de Cito Intelligentietest VO gaat het vooral om het geven van ondersteunende informatie bij vragen die de plaatsing van de leerling

in een schooltype of –niveau betreffen. In dit opzicht is analyse van de relatie met diverse criteria relevant te noemen. In ons geval komen de volgende criteria in aanmerking:

- Niveau-inschatting van de mentor op de school voor voortgezet onderwijs
- Doorstroomadvies van de leerkracht bij de overgang van basisonderwijs naar voortgezet onderwijs
- Het prestatieniveau van de leerlingen zoals bepaald door middel van de Eindtoets Basisonderwijs
- Toetsscores op gestandaardiseerde toetsen, afgenomen in het voortgezet onderwijs
- De feitelijke plaatsing en doorstroom van leerlingen binnen het voortgezet onderwijs.

In het navolgende zullen we resultaten van analyses met betrekking tot de genoemde criteria beschrijven.

Niveau-inschatting van de mentor op de school voor voortgezet onderwijs

Op een deel van de scholen die deelnamen aan het normeringsonderzoek is aan de mentor gevraagd om een niveau-indicatie voor zijn of haar leerlingen af te geven. Een beperkt aantal mentoren was hiertoe bereid. Het betreft dezelfde scholen als de scholen die gegevens beschikbaar stelden uit het Cito Volgsysteem voortgezet onderwijs (zie paragraaf 6.1). Dit leverde voor 221 leerlingen een niveau-indicatie op. Zoals eerder aangegeven ging het om een groep van hoofdzakelijk havo- en vwo-leerlingen. Slechts bij drie leerlingen gaf de mentor een niveau onder havo-niveau aan. Tabel 6.8 geeft een overzicht van de resultaten; hierbij zijn de drie leerlingen met een lager niveau dan havo buiten beschouwing gelaten. Bij de andere niveaus zijn de leerlingaantallen voldoende groot om een duidelijk beeld te krijgen.

Tabel 6.8 Inschatting niveau door mentor: gemiddelde IQ-waarden en toetsscores van het Leerling- en adviessysteem VO

	havo			*havo/vwo*			*vwo*		
Score	*Gem.*	*SD*	*N*	*Gem.*	*SD*	*N*	*Gem.*	*SD*	*N*
Leerjaar-IQ	110,7	10,10	66	115,3	8,92	32	117,8	9,47	113
Leeftijd-IQ	111,6	9,75	69	116,6	8,47	35	118,5	9,25	114
Ned. leesvaardigheid	240,6	20,60	75	250,9	12,31	36	255,9	23,41	119
Ned. woordenschat	260,2	31,15	73	289,0	36,69	36	288,7	47,11	119
Rekenen-wiskunde	250,3	18,29	74	261,2	19,38	36	267,2	23,30	119
Studievaardigheden	244,2	22,73	73	254,7	18,75	36	248,4	26,67	25
Engels	243,8	25,25	74	253,6	27,08	36	263,2	25,00	25

*) Significant p<0,001

De gemiddelde IQ-scores voor de (geschatte) niveaugroepen bevestigen de verwachtingen: havo-leerlingen scoren het laagst, vwo-leerlingen het hoogst; de groep havo/vwo-leerlingen neemt een tussenpositie in. Aan de gemiddelde toetsscores die aan de tabel zijn toegevoegd is te zien dat de niveau-inschatting van de mentor hout snijdt. De gemiddelden zijn in overeenstemming met het aangegeven niveau, met een uitzondering voor de lage score op studievaardigheden voor de groep vwo-leerlingen. Variantieanalyses voor leeftijd-IQ ($F_{2, 211}$ = 11,969; p<0,001) en leerjaar-IQ ($F_{2, 217}$ = 11,263; p<0,001) laten zien dat de gemiddelden duidelijk en significant van elkaar verschillen.

Doorstroomadvies van de leerkracht bij de overgang van basisonderwijs naar voortgezet onderwijs

Voor de leerlingen die in mei-juni 2009 aan het onderzoek naar soortgenootvaliditeit hebben deelgenomen (zie paragraaf 6.1) zijn ook de gegevens van de Eindtoets Basisonderwijs beschikbaar. In het kader van de reguliere afname van de Eindtoets wordt ook geregistreerd wat het advies is van een leerkracht ten aanzien van het onderwijsniveau van elke leerling. In tabel 6.9 is voor 515 leerlingen aangegeven welk doorstroom-advies er over hen is uitgebracht. Daarnaast is per adviescategorie het gemiddelde IQ aangegeven.

De tabel maakt duidelijk dat de gemiddelde IQ's keurig en zoals verwacht oplopen met het niveau van de adviescategorieën. Een enkelvoudige variantieanalyse liet zien dat de gemiddelden *overall* significant van elkaar verschillen ($F_{8, 514} = 65{,}833$; $p < 0{,}001$).

Tabel 6.9 Doorstroomadvies leerkracht basisonderwijs: gemiddelde IQ-scores en standaarddeviatie voor de onderscheiden adviescategorieën

Doorstroomadvies leerkracht	*N*	*%*	*Leeftijd-IQ*	
			M	*SD*
bb	18	3,5%	88,9	10,67
bb/kb	37	7,2%	89,1	8,98
kb	48	9,3%	96,7	9,16
gt	122	23,7%	100,6	9,11
gt/havo	60	11,7%	102,5	9,21
gt/havo/vwo	12	2,3%	107,9	10,83
havo	63	12,2%	109,1	9,64
havo/vwo	87	16,9%	114,4	9,74
vwo	68	13,2%	120,9	8,78
totaal	515	100%		

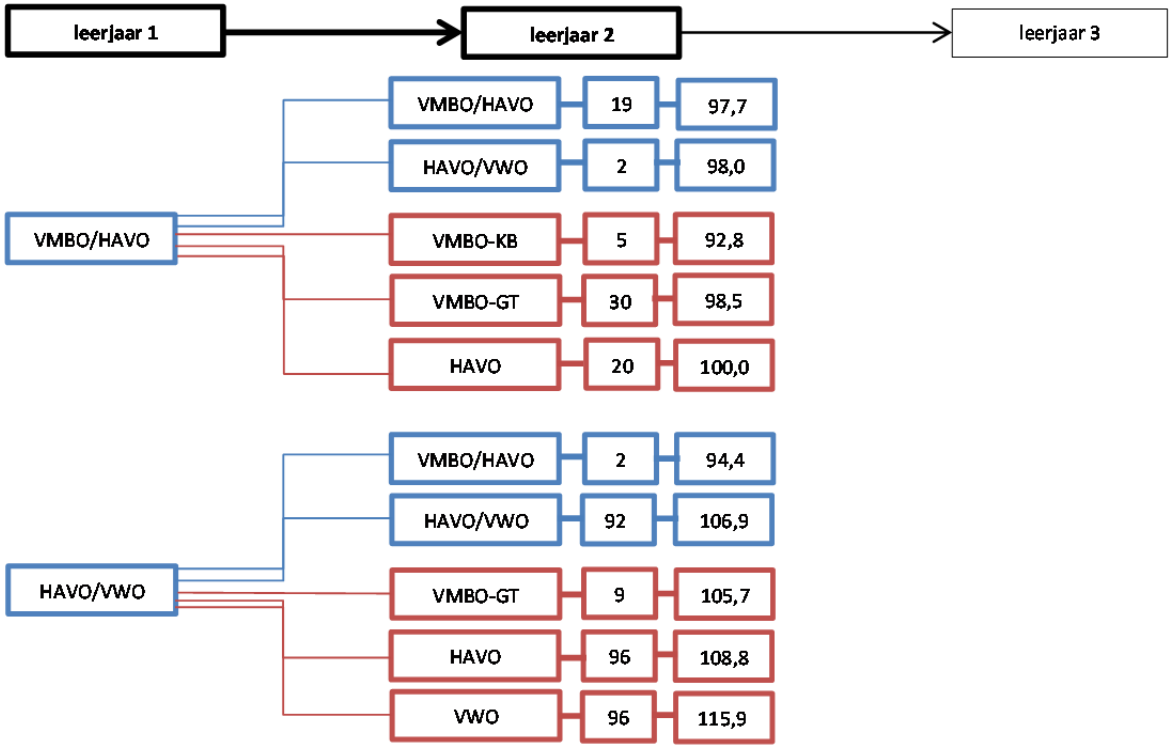
Het prestatieniveau van de leerlingen zoals bepaald door middel van de Eindtoets Basisonderwijs
In paragraaf 6.1 hebben we al laten zien (in tabel 6.4) dat de correlatie van de Intelligentietest VO met de score op de Eindtoets Basisonderwijs om en nabij 0,75 bedraagt. Om precies te zijn, 0,74 voor basisschoolleerlingen (leeftijd-IQ) en 0,76 voor leerlingen van het VO (leerjaar-IQ). Dat is dus een wat lagere samenhang met de Eindtoets dan de NIO (0,78) laat zien. Deze lagere samenhang is volledig in overeenstemming met de uitgangspunten van de testconstructeurs. Het was immers de bedoeling om de Intelligentietest VO zo zuiver mogelijk samen te stellen uit taken die zo weinig mogelijk van doen hebben met wat er op school aan kennis en vaardigheid wordt verworven (zie hoofdstuk 1). Op basis van dit uitgangspunt is het mogelijk om een zo groot mogelijk contrast te creëren met leervorderingentoetsen. Juist de combinatie van de Intelligentietest VO en leervorderingentoetsen (zoals bijvoorbeeld die uit het Cito Volgsysteem voortgezet onderwijs) geeft optimale informatie over de vraag wat "er in zit" en "wat er uit komt".

Toetsscores op gestandaardiseerde toetsen, afgenomen in het voortgezet onderwijs
Ook de samenhang tussen intelligentie-indicatoren met scores op de toetsen uit het Cito Volgsysteem voortgezet onderwijs is in de vorige paragraaf al aan de orde geweest bij de bespreking van de convergente versus discriminante validiteit van de Intelligentietest VO (zie tabel 6.4). De correlaties tussen leeftijd-IQ en leerjaar-IQ aan de ene kant en de toetsscores aan de andere kant herhalen we hier nog eens. Ze liggen tussen 0,28 voor Engels en 0,64 voor Rekenen-wiskunde. Deze samenhangen lijken aan de lage kant, maar we moeten in aanmerking nemen dat we helaas alleen konden beschikken over een selecte groep leerlingen van havo/vwo-niveau, met een gemiddeld IQ groter dan 111 (vergelijk tabel 6.4), dus met een grote 'restriction of range'. Het is moeilijk te zeggen wat de invloed van deze restrictie is op de correlaties. Op grond van vergelijking van de correlaties tussen domeinscores en IQ voor deze selecte groep met dezelfde correlaties voor de volledige normgroep, schatten we in dat de correlaties met waarden tussen 0,05 en 0,15 naar boven moeten worden bijgesteld. Ook dienen we het feit mede in aanmerking te nemen dat correlaties tussen algemene intelligentie en specifieke toetsscores aanzienlijk lager zijn dan de correlatie tussen algemene intelligentie en een algemene leervorderingenscore zoals de totaalscore op de Eindtoets.

De feitelijke doorstroom van leerlingen in het voortgezet onderwijs

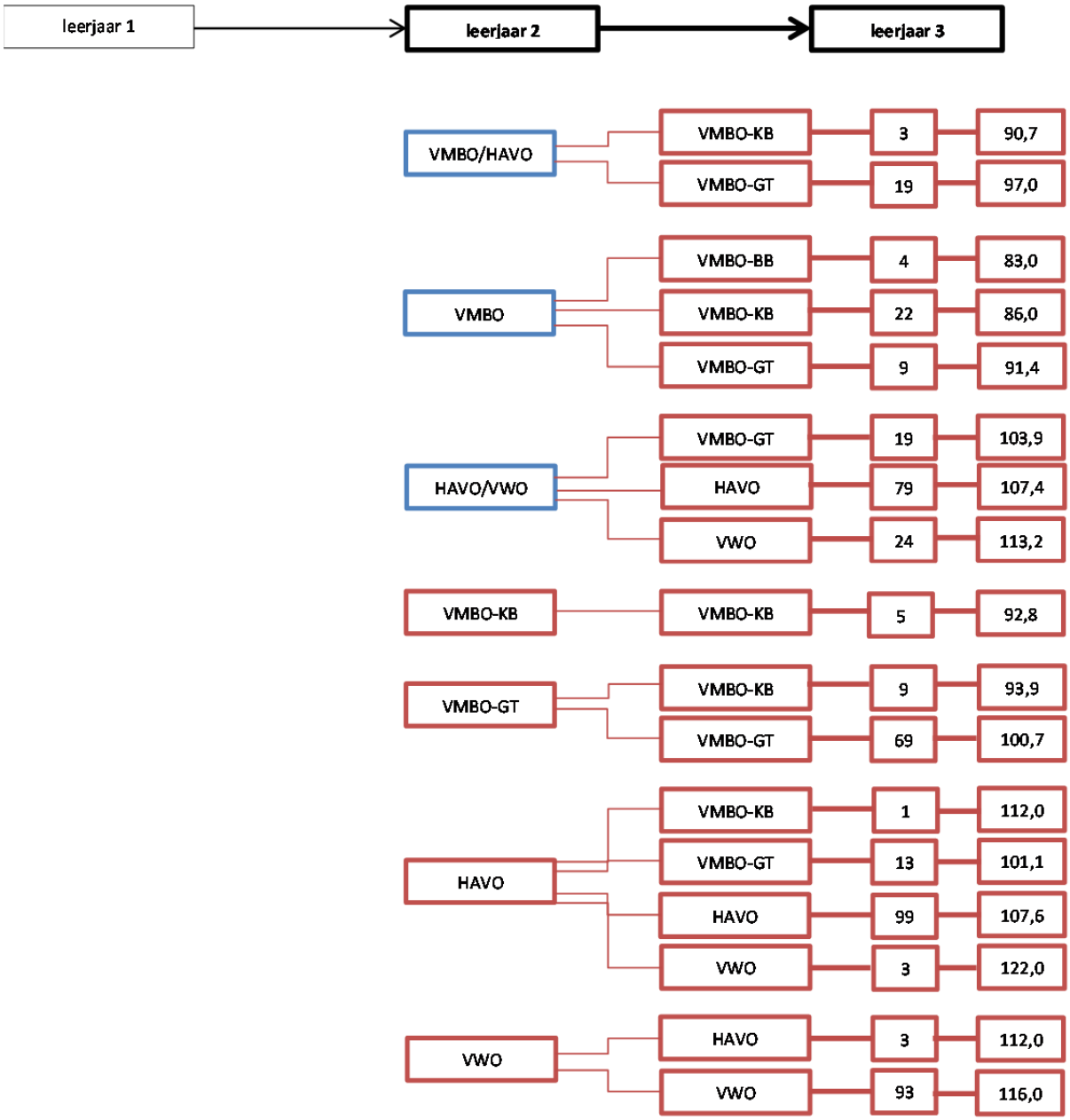
Een belangrijke functie van de Intelligentietest VO is behulpzaam te zijn bij de plaatsing van leerlingen in de onderbouw van het voortgezet onderwijs op momenten dat er keuzes gemaakt moeten worden. Dat is bijvoorbeeld het geval als leerlingen vanuit hun plaats in een bepaald brugjaar een keuze moeten maken uit de onderwijsniveaus die aansluiten bij dat brugjaar. Of in situaties waarin een leerling al geplaatst is en de prestaties in een schooltype tegenvallen, zodat heroverweging van een keuze aan de orde is.

Figuur 6.5 Feitelijke doorstroom binnen het voortgezet onderwijs: van leerjaar 1 naar leerjaar 2



Bij een test die bij dit soort keuzes functioneel moet zijn, mag verwacht worden dat de gemiddelde scores voor bepaalde categorieën (schooltypen) een beeld vertonen dat aansluit bij het niveau. Als bijvoorbeeld een groep leerlingen uitstroomt uit een brugjaar havo/vwo, dienen de leerlingen die in het vwo geplaatst worden gemiddeld een hogere testscore te hebben dan de leerlingen die doorstromen naar het havo.
We hebben vier scholen bereid gevonden om voor in totaal 637 leerlingen door te geven op welke wijze deze leerlingen feitelijk zijn doorgestroomd in het voortgezet onderwijs van leerjaar 1 naar leerjaar 2 en het jaar daarop van leerjaar 2 naar leerjaar 3. Niet alle scholen hebben voor alle overgangen (leerjaar 1 naar leerjaar 2, leerjaar 2 naar leerjaar 3) gegevens verstrekt. Bovendien ontbreken soms gegevens, bijvoorbeeld omdat een leerling vanwege verhuizing de school heeft verlaten. In figuur 6.5 hebben we de gemiddelde scores gegeven bij de overgang van leerjaar 1 naar leerjaar 2. Daarbij zijn brugklastypen blauw en de overige (categoriale) schooltypen in rood aangeduid. Zoals te zien is vertonen alle gemiddelden een verwacht patroon. Zo resulteert bijvoorbeeld de doorstroom uit de eerstejaars brugklas havo/vwo naar brugklassen vmbo/havo en havo/vwo in duidelijk verschillende gemiddelden voor de betreffende leerlingen (94,4 versus 106,9). Hetzelfde geldt voor de uitstroom naar de categoriale schooltypen havo en vwo (108,8 versus 115,9), terwijl de afstroom naar vmbo-gt een lager gemiddelde van 105,7 laat zien.

Figuur 6.6 Feitelijke doorstroom binnen het voortgezet onderwijs: van leerjaar 2 naar leerjaar 3



De situatie bij de overgang van leerjaar 2 naar leerjaar 3 is aanzienlijk complexer vanwege het grotere aantal verschillende schooltypen. Verder zijn ook de gegevens van een school in deze analyses betrokken met een brugklas vmbo in leerjaar 2. Te zien is dat uiteindelijk alle leerlingen in leerjaar 3 zijn geplaatst in een categoriaal schooltype. Steeds vertonen per onderscheiden onderwijsniveau de gemiddelden in de uitstroomcategorieën de verwachte verschillen in hoogte. Dat is ook het geval als er sprake is van afstroom naar een lager schooltype. Er is slechts één uitzondering: een uit het havo naar vmbo-kb uitstromende leerling met een leerjaar-IQ van 112,2 dat men eerder bij havo dan vmbo-kb zou verwachten. Deze gevallen kunnen voorkomen, bijvoorbeeld wanneer een leerling liever kiest voor een specifieke beroepsopleiding in plaats van een algemeen vormende opleiding, of in het geval van ziekte of sociaal-emotionele problematiek.

Tabel 6.10 Gemiddeld leerjaar-IQ per brugklas- en schooltype voor leerjaar 1 tot en met 4 VO

Leerjaar 1	*Gemiddelde*	*SD*	*N*	*Anova*
Brugjaar vmbo	83,1	9,4	22	$F_{2,394} = 101,742$
Brugjaar vmbo/havo	98,3	8,8	78	P < 0,001
Brugjaar havo/vwo	110,3	10,7	295	

Leerjaar 2	*Gemiddelde*	*SD*	*N*	*Anova*
Brugjaar vmbo	87,2	8,7	36	$F_{6,475} = 47,800$
Brugjaar vmbo/havo	96,1	9,6	22	P < 0,001
Brugjaar havo/vwo	108,0	11,3	122	
vmbo-kb	92,8	9,8	5	
vmbo-gt	99,8	9,2	79	
havo	107,3	10,8	116	
vwo	115,9	8,4	96	

Leerjaar 3	*Gemiddelde*	*SD*	*N*	*Anova*
vmbo-bb	83,0	6,1	4	$F_{4, 536} = 67,872$
vmbo-kb	90,4	10,8	44	P < 0,001
vmbo-gt	99,2	12,6	188	
havo	107,6	10,3	181	
vwo	115,6	8,8	120	

Leerjaar 4	*Gemiddelde*	*SD*	*N*	*Anova*
vmbo-kb	96,2	9,1	13	$F_{1,96} = 0,180$
vmbo-gt	98,0	14,9	84	ns

In tabel 6.10 zijn nog eens alle gemiddelden in leerjaar-IQ voor de verschillende brugjaren en schooltypen bij elkaar gezet. Dit hebben we voor alle leerjaren afzonderlijk gedaan. Bij de school met een brugjaar vmbo in leerjaar 2 was voor een deel van de leerlingen bekend dat deze ook in leerjaar 1 in een brugjaar vmbo zaten; deze leerlingen zijn aan de tabel toegevoegd. Van een beperkt deel van de leerlingen was ook bekend hoe zij doorstroomden naar leerjaar 4. Ook deze gegevens zijn aan tabel 6.10 toegevoegd.
De tabel geeft een goede en door de grotere aantallen ook consistentere indruk van de gemiddelde leerjaar-IQ-scores die bij elk schooltype passen. Het is duidelijk dat de gemiddelden steeds keurig volgens verwachting oplopen met het onderwijsniveau. Bij de resultaten voor leerjaar 4 past nog de kanttekening dat het verschil tussen de groep vmbo-kb- en de groep vmbo-gt-leerlingen gering is en niet significant. Nadere inspectie van de gegevens liet zien dat het hier om een kleine groep leerlingen gaat van één school, waarbij negen van de dertien leerlingen in een eerder leerjaar waren afgestroomd van vmbo-gt naar vmbo-kb.

7 Rapportage en interpretatie

7.1 Toelichting leerlingrapport

Leerjaar-IQ
Bij de leerjaar-IQ-score wordt een vergelijking gemaakt met leerjaargenoten. De score loopt uiteen van 70 tot en met 130. De gemiddelde score is 100.
Wanneer de score van de Intelligentietest gebruikt wordt om te adviseren bij de keuze van een onderwijsniveau dan geeft het leerjaar-IQ zinvolle informatie. Er wordt een vergelijking gemaakt met leerlingen die zich in hetzelfde leerjaar bevinden.

Leeftijd-IQ
Bij leeftijd-IQ-score wordt een vergelijking gemaakt met leeftijdgenoten. De score loopt uiteen van 70 tot en met 130. Bij een intelligentietest is het gebruikelijk een leeftijdnormering toe te passen. De leerling wordt dan vergeleken met leerlingen die even oud zijn.
Over het algemeen zal de leerjaar-IQ-score van een leerling niet veel verschillen van zijn leeftijd-IQ-score. Er is een correlatie van 0,99 tussen leeftijd- en leerjaar-IQ. Alleen wanneer een leerling versneld of vertraagd is (is blijven zitten of een klas heeft overgeslagen) kunnen de scores verder uit elkaar liggen.

Het 80%-betrouwbaarheidsinterval
Het 80%-betrouwbaarheidsinterval geeft aan in welk gebied de leerling zal scoren bij een herhaalde afname. Iedere test heeft een zekere mate van onnauwkeurigheid, zodat het resultaat bij de tweede afname enigszins kan afwijken van het resultaat bij eerste afname. Het betrouwbaarheidsinterval is gebaseerd op de lokaal geschatte betrouwbaarheid (zie hoofdstuk 5). Elke IQ-score heeft dan ook zijn eigen betrouwbaarheidsinterval. Het betrouwbaarheidsinterval geeft een belangrijk surplus aan informatie vergeleken met de concreet vastgestelde IQ-score. Het geeft aan met welke nauwkeurigheid men een uitspraak kan doen over de intelligentie van een leerling.

Percentielscore
Op het leerlingrapport wordt de leerjaar-IQ-score en ook de leeftijd-IQ-score omgezet naar een percentielscore. De percentielscore geeft aan welke positie de leerling met zijn testresultaat inneemt in vergelijking met leerjaargenoten en leeftijdgenoten. Een percentielscore van bijvoorbeeld 25 betekent dat 25 procent van de leerlingen dezelfde score of een lagere score heeft behaald. Een percentielscore van 80 betekent dat 80 procent van de leerlingen dezelfde of een lagere score heeft gehaald; 20 procent van de leerlingen heeft dan een beter testresultaat behaald. De maximaal te behalen percentielscore bij zowel leerjaar als leeftijd is 100.

Percentielscores per domein
De percentielscores voor Figuren, Woorden en Getallen geven aan hoe op de afzonderlijke domeinen gescoord is. Het is zinvol deze domeinen onderling te vergelijken en te bepalen of de leerling op alle domeinen ongeveer gelijk gescoord heeft. Het scorebereik van de percentielscores per domein loopt van 1 tot 100.

Grafiek: vergelijking leerjaar-IQ score met onderwijsniveau
In de grafiek wordt de score van de leerling vergeleken met de scores van leerlingen van de verschillende onderwijsniveaus. De verticale streep is de leerjaar-IQ score van de leerling.
De balkjes naast de onderwijsniveaus geven aan waar de meeste scores van de leerlingen van elk onderwijsniveau liggen (zie toelichting in hoofdstuk 4). Het rondje op de balkjes is de middelste score van elk onderwijsniveau (de mediaan).

Het onderwijsniveau waarvan het rondje het dichtst bij de scorelijn van de leerling ligt, is het onderwijsniveau dat het meest in aanmerking komt voor een leerling op basis van zijn intelligentiescore.

Specifieker gezegd: In de grafiek wordt de leerjaar-IQ-score van de leerling afgezet tegen de score-verdelingen van de verschillende onderwijsniveaus. De verticale streep geeft de leerjaar-IQ-score van de leerling weer. De balkjes geven de scoreverdeling aan per onderwijsniveau waarbij telkens de bovenste en onderste 20 procent van de scores zijn weggelaten. Het rondje in de scoreverdeling geeft de mediaan van elk onderwijsniveau aan. De scoreverdelingen van de onderwijsniveaus zijn gebaseerd op de IQ-scoreverdelingen van de leerlingen uit leerjaar 3 omdat dat leerjaar het beste weergeeft waar leerlingen met een bepaalde IQ-score uiteindelijk terechtkomen. Meer informatie over verschillen in gemiddelden naar onderwijsniveau is te vinden in hoofdstuk 6 (paragraaf 6.2 over criteriumvaliditeit).

Grenswaarden en mediaan per onderwijsniveau

	ondergrens	*mediaan*	*bovengrens*
VWO	107	117	125
HAV	99	105	113
GT	95	102	109
KB	85	92	98
BB	77	84	90
BB+	70	78	86

Afkortingen

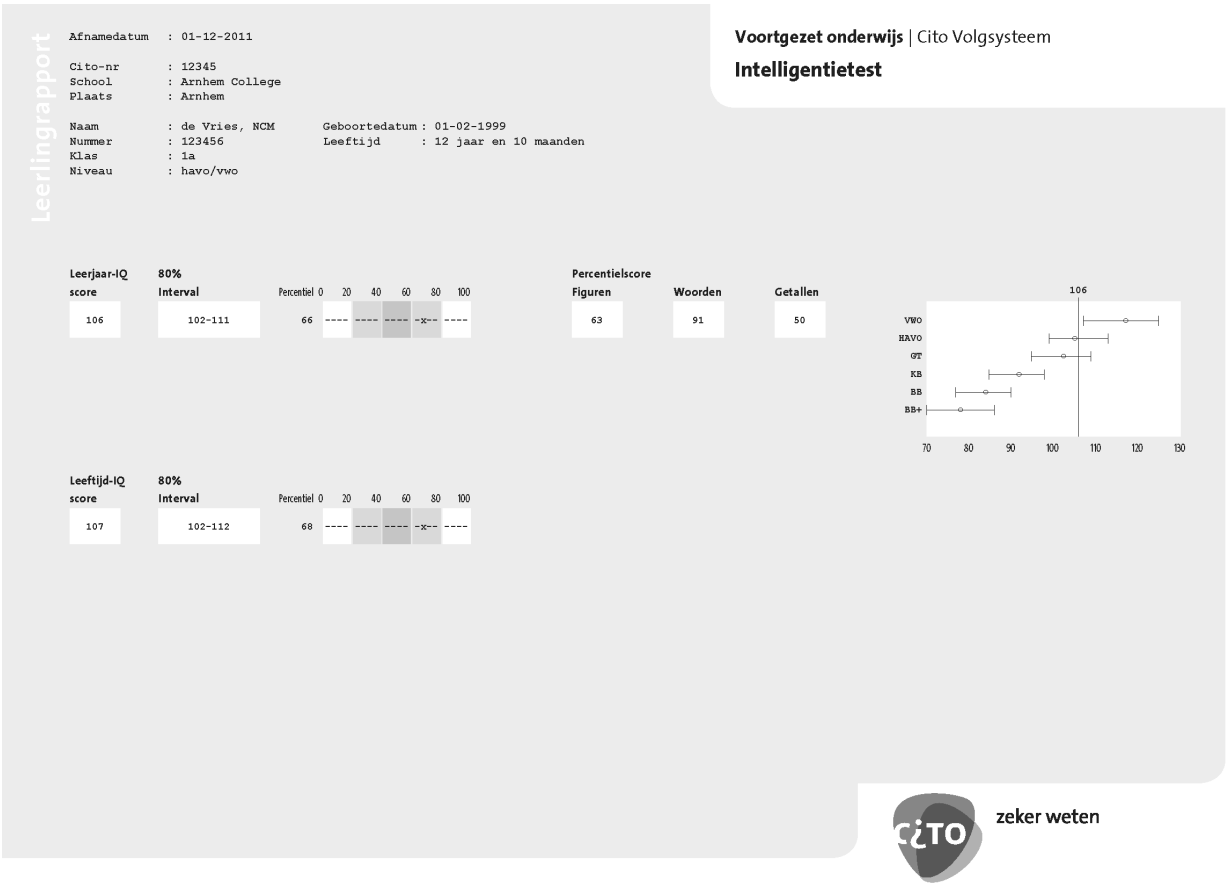
Dit zijn de afkortingen voor onderwijsniveaus en leerwegen:

VWO	vwo
HAV	havo
GT	vmbo- gemengde en theoretische leerweg
KB	vmbo - kaderberoepsgerichte leerweg
BB	vmbo - basisberoepsgerichte leerweg zonder lwoo
BB+	vmbo - basisberoepsgerichte leerweg met lwoo

7.2 Voorbeelden van een leerlingrapport

We illustreren de betekenis van de scores op het leerlingrapport met twee voorbeelden.

Voorbeeldrapportage1



Leerlingrapport

Afnamedatum : 01-12-2011

Cito-nr : 12345
School : Arnhem College
Plaats : Arnhem

Naam : de Vries, NCM Geboortedatum : 01-02-1999
Nummer : 123456 Leeftijd : 12 jaar en 10 maanden
Klas : 1a
Niveau : havo/vwo

Voortgezet onderwijs | Cito Volgsysteem
Intelligentietest

Leerjaar-IQ score	80% Interval	Percentiel 0 20 40 60 80 100
106	102-111	66 ---- ---- ---- -x-- ----

Leeftijd-IQ score	80% Interval	Percentiel 0 20 40 60 80 100
107	102-112	68 ---- ---- ---- -x-- ----

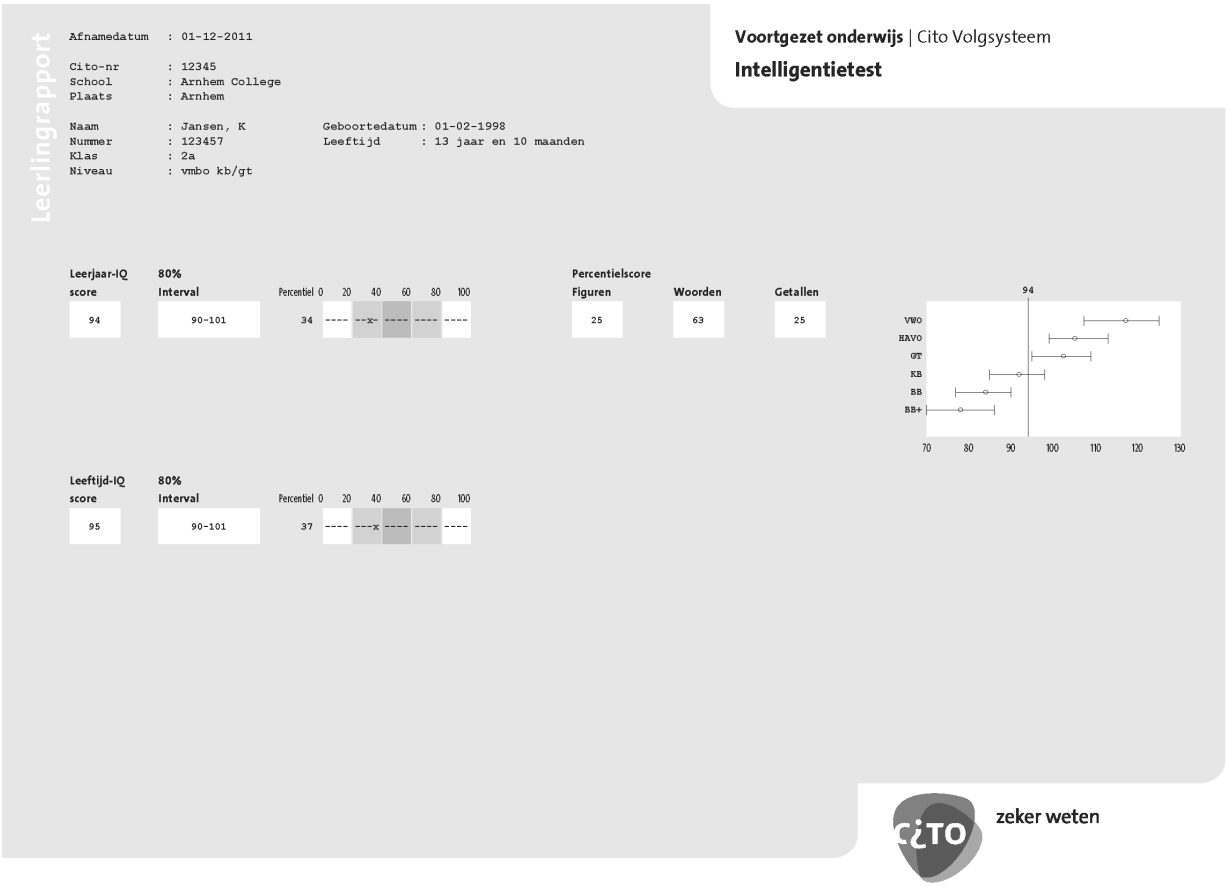
Percentielscore Figuren	Woorden	Getallen
63	91	50

zeker weten

Toelichting op het rapport van NCM de Vries (Nicole)

In de vergelijking met leerjaargenoten behaalt Nicole een IQ-score van 106. In de vergelijking met leeftijd-genoten haalt ze een IQ-score van 107. De 80%-betrouwbaarheidsscore geeft aan dat de verwachte score bij herhaalde afname tussen de 102 en 111 zal liggen bij een leerjaarvergelijking (i.e. leerjaar-IQ) en tussen de 102 en 112 bij een leeftijdvergelijking (i.e. leeftijd-IQ).
Nicole haalt een percentielscore van 66 bij een leerjaarvergelijking en van 68 bij een leeftijdvergelijking. Dat houdt in dat Nicole een score haalt die hoger is dan (of gelijk is aan) 66 procent van de leerjaar-populatie (Nicole zit in leerjaar 1) en 68 procent van de leeftijdpopulatie (Nicole is 12 jaar). 34 procent van de leerjaarpopulatie haalt een hogere score dan Nicole.
De percentielscore voor het onderdeel Figuren is 63, voor het onderdeel Woorden 91 en het onderdeel Getallen 50.
Nicole heeft op alle onderdelen een score behaald die behoort tot minimaal de hoogste 50%. De laagste score haalt ze op het onderdeel Getallen en de hoogste op het onderdeel Woorden. De scores op de verschillende onderdelen lopen behoorlijk uiteen. Bij Getallen heeft ze een doorsnee percentielscore gehaald (percentiel 50) terwijl ze bij het onderdeel Woorden een percentielscore haalt van 91.
Uit de grafiek valt op te maken dat de leerjaar-IQ-score van Nicole net boven het midden van de havo-scores ligt. Dat is te zien aan de rondjes van de scoreverdelingen. Het rondje van havo ligt het dichtst bij de scorelijn van Nicole. Dit betekent dat de prestaties van Nicole op de Intelligentietest VO het best passen bij havo-niveau.

Voorbeeldrapportage 2



Leerlingrapport

Afnamedatum : 01-12-2011

Cito-nr : 12345
School : Arnhem College
Plaats : Arnhem

Naam : Jansen, K
Nummer : 123457
Klas : 2a
Niveau : vmbo kb/gt

Geboortedatum : 01-02-1998
Leeftijd : 13 jaar en 10 maanden

Voortgezet onderwijs | Cito Volgsysteem
Intelligentietest

Leerjaar-IQ score	80% Interval	Percentiel 0 20 40 60 80 100	Percentielscore Figuren	Woorden	Getallen
94	90-101	34 ---- --x- ---- ---- ----	25	63	25

Leeftijd-IQ score	80% Interval	Percentiel 0 20 40 60 80 100
95	90-101	37 ---- ---x ---- ---- ----

zeker weten

Toelichting op het rapport van K Jansen (Koen)

Koen heeft een leerjaar-IQ-score gehaald van 94. In vergelijking met leeftijdgenoten is de IQ-score 95. Het betrouwbaarheidsinterval geeft aan dat bij herhaalde meting een score wordt verwacht tussen de 90 en 101 zowel bij de leerjaar- als de leeftijd-IQ-score.
De percentielscore bij het leerjaar-IQ is 34 en bij het leeftijd-IQ 37. Bij het leerjaar-IQ scoort Koen beter dan of even hoog als 34 procent van de leerjaargenoten. Dat houdt in dat 66 procent van de leerjaargenoten een hogere score behaalt dan Koen.

De percentielscores voor Figuren en Getallen zijn identiek, percentiel 25. De hoogste score behaalt Koen bij het onderdeel Woorden. Met een percentielscore van 63 ligt deze net boven de gemiddelde score van de normgroep.

De grafiek laat zien dat de score van Koen boven de middelste score ligt van het vmbo-kb. Wanneer zijn score wordt vergeleken met de score van het vmbo-gt dan wordt duidelijk dat de score van Koen net onder percentiel 20 van dat onderwijsniveau ligt. De scorelijn van Koen ligt uiteindelijk het dichtst bij het rondje van het onderwijsniveau vmbo-kb.

7.3 Overige rapportages

Naast het leerlingrapport zijn er aanvullende rapportages beschikbaar in de *Internetrapportage*.
De *Internetrapportage* is een online rapportageservice van het Cito Volgsysteem voortgezet onderwijs. Dit programma biedt de mogelijkheid de leerlingrapportages online op te vragen. Daarnaast kunnen er met dit programma groepsoverzichten gemaakt worden.

De volgende rapportages kunnen opgevraagd worden:

- rapportage leerling;
- groepsoverzicht.

De *Internetrapportage* is te vinden op http://portal.cito.nl. Met behulp van de aan de school toegekende inloggegevens kan men in deze omgeving de resultaten van de leerlingen inzien. In de handleiding Internetrapportage staat een uitgebreide toelichting op de verschillende rapportages. Deze handleiding is te vinden op de portal en op www.cito.nl.

8 Samenvatting en conclusies

De Cito Intelligentietest VO is een algemene intelligentietest die bedoeld is voor leerlingen in leerjaar 1, 2 en 3 van het voortgezet onderwijs. De test kan worden ingezet bij vragen die betrekking hebben op het meest geschikte onderwijsniveau voor een leerling op basis van diens cognitieve capaciteiten. De test gaat uit van definities van algemene intelligentie zoals deze voorgesteld zijn door onder andere Resing en Drenth (2001) en Wechsler (1944). Op basis van deze definities probeert de test zicht te krijgen op de intellectuele vermogens door afname van cognitieve taken van verschillende aard en met betrekking tot verschillende domeinen: getallen, woorden en figuren. Bij alle taken staan de basale redeneervaardigheden en de hogere mentale processen voorop zoals deze met name naar voren komen als aspecten van *Fluid Intelligence*, de brede tweede orde factor die volgens Carroll (1993) in hoge mate de algemene intelligentie bepaalt. Omdat de vaardigheden en processen die bij deze taken een rol spelen in relatief geringe mate het resultaat zijn van kennisverwerving en (leer)ervaring, kan een optimaal contrast worden gecreëerd tussen capaciteiten (“wat zit er in?”) en leervorderingen (“wat komt er uit?”). Op basis van de testscore kan zowel een leerjaar- als een leeftijd-IQ (laatstgenoemde alleen voor 11- tot en met 14-jarigen) worden berekend. Meer over de theoretische achtergrond van de test, meetpretentie, functie en doelgroep van de test is te lezen in hoofdstuk 2.
In hoofdstuk 3 kwamen de zes verschillende taken in de intelligentietest (Figuurclassificatie, Figuurmatrix, Woordclassificatie, Woordanalogie, Getalreeks en Getalanalogie) aan de orde. De meetpretentie en opdracht bij elke taak werden aan de hand van een voorbeeld besproken. Verder doet dit hoofdstuk verslag van de verschillende vooronderzoeken die aan de constructie van de test ten grondslag lagen.
Het hoofdstuk wordt afgesloten met de bespreking van de belangrijkste kenmerken van de geselecteerde items.
Hoofdstuk 4 handelt over de kalibratie en normering van de test. Omdat gebruikgemaakt is van – bij de constructie van intelligentietests vrij ongebruikelijke – IRT-modellen is in dit hoofdstuk een uitgebreide toelichting gegeven op de gehanteerde procedures en het gehanteerde meetmodel (het mariginale One Parameter Logistic Model – OPLM). Dat geldt ook voor de op IRT gebaseerde, Bayesiaanse methoden om de steekproef te wegen en de betrouwbaarheden te schatten. Het is niet eenvoudig om een adequate representatieve steekproef te vormen wanneer men (a) een collectieve, schoolgerichte benadering kiest voor de dataverzameling en (b) naast een leerjaarnormering ook een normering naar leeftijd wil realiseren. In hoofdstuk 4 is uitgebreid ingegaan op de wijze waarop populatieverdelingen voor leeftijd, leerjaar en schooltype werden gegenereerd. We rapporteerden hoe we op basis van een steekproefkader scholen hebben geworven, waarbij het streven was om representativiteit naar regio en mate van verstedelijking op schoolniveau te realiseren. In tweede instantie werd er gewogen naar leerjaar, leeftijd en schooltype om de steekproef voor deze variabelen op leerlingniveau representatief te maken. Daarnaast geeft dit hoofdstuk informatie over de kalibratie en over de verdelingskenmerken van de belangrijkste testscores.
Ook de betrouwbaarheden van de test zijn via methoden uit de moderne testtheorie vastgesteld. Hoofdstuk 5 doet hiervan verslag. De betrouwbaarheid van het leeftijd- en leerjaar-IQ bedraagt in alle normgroepen 0,95 en voldoet daarmee dus ruimschoots aan de eisen die gesteld worden aan tests op basis waarvan belangrijke beslissingen kunnen worden genomen. Voor de domeinscores voor Getallen, Woorden en Figuren liggen de betrouwbaarheden tussen 0,83 en 0,95, waarmee zij eveneens ruimschoots voldoen aan de eisen die aan dit soort deelscores worden gesteld. In aanvulling hierop is test-hertestonderzoek uitgevoerd. De berekende test-hertestbetrouwbaarheden liggen gemiddeld omstreeks 0,10 lager dan de Bayesiaans geschatte betrouwbaarheden (wat voor dit type waarden gebruikelijk is) en bevestigen daarmee de hoge betrouwbaarheid van de test. Op basis van de gebruikte procedures kan niet één betrouwbaarheidsinterval voor de gehele testscore worden gegeven omdat de betrouwbaarheidsintervallen lokaal, dus voor elke testscore afzonderlijk zijn geschat. Daarom is grafisch weergegeven hoe de lokale meetnauwkeurigheid van de test er uit ziet.

In hoofdstuk 6 is uitgebreid ingegaan op de begrips- en criteriumvaliditeit van de test.
Ter onderbouwing van de begripsvaliditeit werd op de eerste plaats aandacht besteed aan de kalibratie. Die is voor verschillende leeftijdsgroepen afzonderlijk gebeurd omdat DIF-analyses lieten zien dat dit nodig

was. Ook is de evidentie besproken die erop wijst dat de kalibratie geslaagd is. Daarnaast kwamen de (over het algemeen hoge) intercorrelaties aan de orde en de (hoge) correlaties tussen deeltaakscores en totale testscores. Deze bleken in de verschillende normgroepen steeds hetzelfde patroon te vertonen, wat duidt op invariantie van de teststructuur voor de onderscheiden normgroepen naar leeftijd en leerjaar. Uitgebreid aanvullend onderzoek werd uitgevoerd naar de soortgenootvaliditeit. Correlaties van het leerjaar- en leeftijd-IQ met het NIO leerjaar-IQ bedroegen respectievelijk 0,89 en 0,81. De correlatie tussen leeftijd-IQ en de score op de Cito Intelligentietest EB (een test voor een iets jongere leeftijdsgroep) was met 0,74 iets lager. Uit aanvullend onderzoek naar toetsscores uit het Cito Volgsysteem voortgezet onderwijs bleek dat de intelligentietest het sterkst samenhangt (>0,60) met de toetscores voor Rekenen-Wiskunde. Voor de drie talige toetsscores (Nederlandse leesvaardigheid en woordenschat, Engels) verwachtten we de hoogste correlatie met het domein Woorden. Voor Rekenen-Wiskunde daarentegen verwachtten we de hoogste correlatie met Figuren en Getallen. Alle verwachtingen werden bevestigd door de data en zijn daarmee dus ondersteunend voor de divergente validiteit. Er werden diverse analyses uitgevoerd met betrekking tot verschillen tussen subgroepen. Deze lieten zien dat de verwachte verschillen in gemiddelde optreden voor de onderscheiden leeftijd- en leerjaar(norm)groepen. Hetzelfde geldt voor de verschillende onderwijsniveaugroepen die in het voortgezet onderwijs kunnen worden onderscheiden. Meisjes en jongens bleken ongeveer gelijk te scoren op de test. Dyslectische leerlingen bleken ongeveer even hoog te scoren als niet-dyslectische leerlingen. Daarentegen bleken er duidelijke verschillen naar thuistaal te bestaan: leerlingen die thuis geen Nederlands spreken, maar een andere taal bleken in alle onderwijsniveaugroepen lager te scoren.

Veel aandacht is besteed aan onderzoek naar de criteriumvaliditeit. Op basis van de testscore (i.e. zowel leeftijd- als leerjaar-IQ) zijn leervorderingen zoals gemeten met de Eindtoets Basisonderwijs goed te voorspellen. De correlaties van respectievelijk 0,76 en 0,74 laten zien dat de constructeurs goed geslaagd zijn in hun opzet om door de keuze van “fluïd” redeneertaken voldoende onafhankelijkheid van onderwijs-inhouden te creëren. Er is een duidelijk verband tussen de intelligentietest en het doorstroomadvies dat is afgegeven door de leerkracht op de basisschool bij de overgang naar het voortgezet onderwijs. Hetzelfde geldt voor de niveau-inschatting door de mentor in het voortgezet onderwijs. Onderwijsniveaugroepen met een oplopende moeilijkheidsgraad lieten duidelijke verschillen in (oplopende) gemiddelden zien. Hetzelfde geldt voor de feitelijke plaatsing en doorstroom van leerlingen in de eerste drie leerjaren van het voortgezet onderwijs en de uiteindelijke plaatsing in categoriale schooltypen in leerjaar 3.

Deze wetenschappelijk verantwoording werd afgesloten met een hoofdstuk over scoring en interpretatie van de testresultaten waarin de uitkomsten van de hier gerapporteerde analyses werden betrokken.

Referenties

Bechger, T.M., Maris, G., Verstralen, H.H.F.M., & Beguin, A.A. (2003). Using Classical Test Theory in Combination with Item Response Theory. *Applied Psychological Measurement, 27*, 319-334.

Bechger, T. M., & Maris, G. (2010). *A different view on DIF. RD Rapport 10-04*. Arnhem: Cito.

Boxtel, H.W. van, Snijders, J.Th. & Welten, V.J. (1982). *ISI: Interesse, Schoolvorderingen, Intelligentie. Verantwoording van het prestatiegedeelte & Handleiding voor de gehele testreeks.* Vorm III, publikatie 7. Groningen: Wolters-Noordhoff.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factoranalytic studies*. Cambridge: Cambridge University Press.

Cito (2012). *Cito Volgsysteem voortgezet onderwijs. Intelligentietest. Handleiding.* Arnhem: Cito.

CBS (2011). *CBS Statline.* http://statline.cbs.nl/statweb/

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Dijk, H. van, & Tellegen, P.J. (2004). *Nederlandse Intelligentietest voor Onderwijsniveau. Handleiding en verantwoording. Uitgebreide samenvatting.* Amsterdam: Boom testuitgevers. In 2008 te downloaden van: http://www.testresearch.nl/nio/nioverkort.doc

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests.* Geheel herziene versie mei 2009; gewijzigde herdruk mei 2010. Amsterdam: NIP-COTAN.

Flanagan, D.P. & Kaufman, A.S. (2004). *Essentials of WISC-IV assessment.* New York: Wiley.

Geelhoed, J.W., Struiksma, A.J.C., & Moesker, E.H.M. (2009). Intelligentieonderzoek. In Th. Kievit, J.A. Tak, & J.D. Bosch (red.). *Handboek psychodiagnostiek voor de hulpverlening aan kinderen. Hoofdstuk 12.* Utrecht: de Tijdstroom, 383 – 438.

Hop, M. (2012). *Balans van het rekenen-wiskunde onderwijs halverwege de basisschool 5*. Arnhem: Cito.

Jarque, C. M., & Bera, A.K. (1987). A test for normality of observations and regression residuals, *International* Statistical Review, *55*, 163-172.

Koenker, R. (2005). *Quantile Regression (Econometric Society Monographs)*. New York: Cambridge University Press.

Kort, W., Schittekatte, M., Dekker, P.H., Verhaeghe, P., Compaan, E.L., Bosmans, M. & Vermeir, G. (2005). *WISC-III-NL. Wechsler Intelligence Scale for Children – III. David Wechsler. Handleiding en verantwoording.* Amsterdam: Harcourt, NIP Dienstencentrum.

Maas, H.L.J. van der, Dolan, C.V., Grasman, R.P.P.P., Wicherts, J.M., Huizenga, H.M. & Raaijmakers, M.E.J. (2006). A dynamic model of general Intelligence: the positive manifold of intelligence by mutualism. *Psychological Review, 113 (4*), 842-861.

Maris, G., & Bechger, T. M. (2009). *Minorization Algorithms for the Rasch model. RD Report 09-03*. Arnhem: Cito.

Marsman, M., Maris, G., Bechger, T.M., & Glas, C.A.W. (2011). *A conditional composition algorithm for latent regression. RD Rapport 11-02*. Arnhem: Cito.

Marsman, M., Maris, G., Bechger, T.M., & Glas, C.A.W. (In voorbereiding). *A non-parametric estimator of latent variable distributions*.

Mislevy R., Johnson E., & Muraki E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics, 17*, 131-154.

Murdoch, D.J. & Chow, E.D. (1996). A graphical display of large correlation matrices. *The American Statistician, 50*, 178-180.

Ramsden, S., Richardson, F.M., Josse, G. , Thomas, M.S., Ellis, C., Shakeshaft, C., Seghier, M.L., & Price, C.J. (2011). Verbal and non-verbal intelligence changes in the teenage brain. *Nature, 479,* 113-116.

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

Resing, W.C.M. & Drenth, P. (2001). *Intelligentie. Meten en weten*. Amsterdam: Uitgeverij Nieuwezijds.

Sternberg, R.J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, *77*(1), 4-20.

Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer-Verlag.

Verhelst, N. D., en Eggen, T. J. H. M. (1989). *Psychometrische en statistische aspecten van peilingonderzoek. PPON-Rapport 4*, 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.

Wechsler, D. (1944). *The measurement of adult intelligence*. Third edition. Baltimore: Williams & Wilkins.

Bijlagen

Tabel B1 Intercorrelaties tussen PVs voor de afzonderlijke deeltaken per leeftijd

Leeftijd 11 (gem, correlatie 0,74)

	pv,FC	pv,FM	pv,GR	pv,GA	pv,WC	pv,WA
pv,FC	1	0,86	0,8	0,8	0,84	0,77
pv,FM	0,86	1	0,7	0,76	0,71	0,69
pv,GR	0,8	0,7	1	0,8	0,64	0,62
pv,GA	0,8	0,76	0,8	1	0,67	0,69
pv,WC	0,84	0,71	0,64	0,67	1	0,82
pv,WA	0,77	0,69	0,62	0,69	0,82	1

Leeftijd 12 (gem, correlatie 0,72)

	pv,FC	pv,FM	pv,GR	pv,GA	pv,WC	pv,WA
pv,FC	1	0,82	0,73	0,7	0,66	0,73
pv,FM	0,82	1	0,73	0,8	0,6	0,7
pv,GR	0,73	0,73	1	0,83	0,58	0,7
pv,GA	0,7	0,8	0,83	1	0,57	0,69
pv,WC	0,66	0,6	0,58	0,57	1	0,9
pv,WA	0,73	0,7	0,7	0,69	0,9	1

Leeftijd 13 (gem, correlatie 0,71)

	pv,FC	pv,FM	pv,GR	pv,GA	pv,WC	pv,WA
pv,FC	1	0,78	0,72	0,71	0,74	0,67
pv,FM	0,78	1	0,7	0,75	0,68	0,69
pv,GR	0,72	0,7	1	0,81	0,62	0,68
pv,GA	0,71	0,75	0,81	1	0,65	0,66
pv,WC	0,74	0,68	0,62	0,65	1	0,84
pv,WA	0,67	0,69	0,68	0,66	0,84	1

Leeftijd 14 (gem, correlatie 0,75)

	pv,FC	pv,FM	pv,GR	pv,GA	pv,WC	pv,WA
pv,FC	1	0,84	0,73	0,7	0,78	0,72
pv,FM	0,84	1	0,74	0,76	0,68	0,76
pv,GR	0,73	0,74	1	0,79	0,75	0,72
pv,GA	0,7	0,76	0,79	1	0,67	0,71
pv,WC	0,78	0,68	0,75	0,67	1	0,88
pv,WA	0,72	0,76	0,72	0,71	0,88	1

Tabel B2 Intercorrelaties tussen PVs voor de afzonderlijke deeltaken per leerjaar

Leerjaar 1 (gem, correlatie 0,71)

	pv,FC	pv,FM	pv,GR	pv,GA	pv,WC	pv,WA
pv,FC	1	0,78	0,7	0,68	0,69	0,67
pv,FM	0,78	1	0,72	0,78	0,64	0,7
pv,GR	0,7	0,72	1	0,82	0,63	0,69
pv,GA	0,68	0,78	0,82	1	0,63	0,67
pv,WC	0,69	0,64	0,63	0,63	1	0,87
pv,WA	0,67	0,7	0,69	0,67	0,87	1

Leerjaar 2 (gem, correlatie 0,73)

	pv,FC	pv,FM	pv,GR	pv,GA	pv,WC	pv,WA
pv,FC	1	0,81	0,73	0,7	0,74	0,69
pv,FM	0,81	1	0,72	0,74	0,67	0,72
pv,GR	0,73	0,72	1	0,79	0,69	0,69
pv,GA	0,7	0,74	0,79	1	0,66	0,66
pv,WC	0,74	0,67	0,69	0,66	1	0,87
pv,WA	0,69	0,72	0,69	0,66	0,87	1

Leerjaar 3 (gem, correlatie 0,76)

	pv,FC	pv,FM	pv,GR	pv,GA	pv,WC	pv,WA
pv,FC	1	0,82	0,72	0,7	0,74	0,73
pv,FM	0,82	1	0,77	0,83	0,72	0,8
pv,GR	0,72	0,77	1	0,79	0,72	0,74
pv,GA	0,7	0,83	0,79	1	0,68	0,76
pv,WC	0,74	0,72	0,72	0,68	1	0,88
pv,WA	0,73	0,8	0,74	0,76	0,88	1

Tabel B3 Leerlingaantallen in basisonderwijs en speciaalonderwijs van schooljaar 2008/2009

Leeftijd	onderwijs	aantal	Totaal per leeftijdjaar
10 jaar	basisonderwijs	184445	
10 jaar	speciaal basisonderwijs	8354	192799
11 jaar	basisonderwijs	176127	
11 jaar	speciaal basisonderwijs	9005	185132
12 jaar	basisonderwijs	70928	
12 jaar	speciaal basisonderwijs	7482	78410
13 jaar	basisonderwijs	9983	
13 jaar	speciaal basisonderwijs	992	10975
14 jaar en ouder	basisonderwijs	312	
14 jaar en ouder	speciaal basisonderwijs	21	333

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

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

Fotografie: Ron Steemers