FACULTEIT PSYCHOLOGIE EN PEDAGOGISCHE WETENSCHAPPEN
DEPARTEMENT PSYCHOLOGIE
ONDERZOEKSGROEP HOGERE COGNITIE EN INDIVIDUELE VERSCHILLEN
CENTRUM VOOR ORGANISATIE- EN PERSONEELSPSYCHOLOGIE
TIENSESTRAAT 102 – 3000 LEUVEN

KATHOLIEKE
UNIVERSITEIT
LEUVEN

Hebben autochtonen en allochtonen gelijke kansen bij selectieproeven met intelligentietests?

Dr. Michel Meulders
Miek Vandenberk
Prof. Dr. Paul De Boeck
Prof. Dr. Karel De Witte
Dr. Rianne Janssen

Maart 2005

Een onderzoek in opdracht van minister Renaat Landuyt, Vlaams minister van Werkgelegenheid en Toerisme, en minister Paul Van Grembergen, Vlaams minister van Binnenlandse Aangelegenheden, in het kader van het VIONA-onderzoeksprogramma.

Met ondersteuning van de administratie Werkgelegenheid en de Dienst Emancipatiezaken.

Dit is het eindrapport voor de doelgroep allochtonen van het VIONA-onderzoeksproject "Psychologische testen en de effecten op instroom van kansengroepen in het Ministerie van de Vlaamse Gemeenschap en in de Vlaamse privébedrijven".

Dit deelproject is een samenwerking tussen de "Onderzoeksgroep Hogere Cognitie en Individuele Verschillen" en het "Centrum voor Organisatie- en Personeelspsychologie" van de Faculteit Psychologie en Pedagogische Wetenschappen van de Katholieke Universiteit Leuven.

Contactadressen:

Dr. Michel Meulders (promotor)
Onderzoeksgroep Hogere Cognitie en Individuele Verschillen
Tiensestraat 102
B – 3000 Leuven

Tel.: 016/32 59 85
E-mail: michel.meulders@psy.kuleuven.ac.be

Miek Vandenberk (onderzoeker)
Onderzoeksgroep Hogere Cognitie en Individuele Verschillen
Tiensestraat 102
B – 3000 Leuven

Tel.: 016/32 59 86
E-mail: miek.vandenberk@psy.kuleuven.ac.be

Inhoudstafel

Hoofdstuk 1: Inleiding

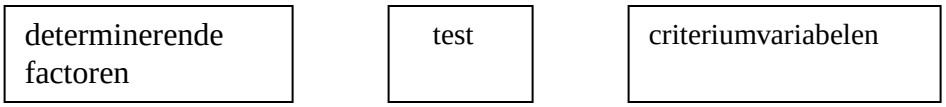
De bedoeling van het project is om te onderzoeken of relevante intelligentietests die gebruikt worden in de context van personeelsselectie een bias vertonen voor elk van drie kansengroepen, namelijk vrouwen, allochtonen en gehandicapten. In dit rapport beschrijven we de resultaten van het onderzoek bij de doelgroep allochtonen. Hiervoor baseren we ons op de analyse van testgegevens die verzameld werden bij VDAB, SELOR en ABL[1.]

Een test of item uit een test vertoont een bias als naast de variabele die men wil meten ook de groep waartoe men behoort het resultaat van een persoon voor een item of voor de gehele test bepaalt. Bij gelijke intelligentie moeten de succeskansen dezelfde zijn, ongeacht de groep waartoe men behoort. De bias kan bestudeerd worden voor individuele items of voor de test in zijn geheel. Veronderstel dat twee personen even intelligent zijn, maar dat de persoon die tot groep A behoort voor een bepaald item een kleinere kans heeft op een juist antwoord dan de persoon die tot groep B hoort. Als dit zich voordoet bij een bepaald item, dan werkt het betreffende item discriminerend in het nadeel van groep A. Het item vertoont dan “itembias”. Het functioneert anders in de onderscheiden subgroepen. In het Engels wordt dit “differential item functioning” of kortweg DIF genoemd. Het item zou dan eigenlijk verwijderd moeten worden als men de twee groepen gelijke kansen wil geven. Als een test een bias vertoont voor verschillende items dan is het interessant om “testbias” (de bias voor de gehele test) of “differential test functioning” (DTF) te onderzoeken. Testbias kan op verschillende manieren geconceptualiseerd worden: (1) In de klassieke testtheorie (KTT) spreekt men van testbias als testscores van verschillende subgroepen een verschillende predictieve validiteit hebben voor het criterium dat men beoogt te meten met de test. (2) In het kader van de itemrespons theorie (IRT) definieert men testbias als het effect van bias in individuele items op de testscores van subgroepen (Shealy en Stout, 1993). Tenslotte merken we op dat in de literatuur de term "bias" soms wordt gereserveerd voor gevallen waarin de constructvaliditeit van de test gewaarborgd is (zie Shealy en Stout, 1993). We zullen in het vervolg van het rapport dit onderscheid niet expliciet maken en de termen DIF en itembias als synoniemen beschouwen voor hetzelfde statistische fenomeen.

[1] We danken de organisaties VDAB, SELOR en ABL voor hun medewerking aan het onderzoek en voor hun bereidwilligheid om de testgegevens voor dit rapport ter beschikking te stellen. Dank aan de verantwoordelijken binnen elke organisatie die instemden om mee te werken aan dit onderzoek en aan de contactpersonen Ann Otte (VDAB), Katrien Brysse (SELOR) en Bert Schreurs (ABL) die steeds zo vriendelijk waren om ons verder te helpen.

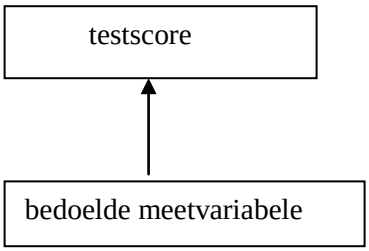
1.1 Conceptueel kader

Het conceptueel kader van de studie van tests kan als volgt beschreven worden:
a. Er wordt een onderscheid gemaakt tussen determinerende factoren, de test en criteriumvariabelen.



De *determinerende factoren* zijn variabelen zoals vooropleiding, vertrouwdheid met de taal, motivatie, de maatschappelijke groep waartoe men behoort (man of vrouw, allochtoon of autochtoon, met of zonder handicap, enz.). De *test* bestaat uit een reeks opgaven of vragen, items genoemd. Doorgaans wordt de som bepaald van de itemscores (bijvoorbeeld de som van het aantal juiste antwoorden) en wordt die "ruwe uitslag" genoemd. Soms wordt die ruwe score omgezet in een afgeleide uitslag op basis van een normering. De *criteriumvariabelen* zijn de variabelen waarin men is geïnteresseerd. Het zijn de variabelen die men wil voorspellen of verklaren. Intelligentietests worden vaak aangewend als predictoren. Ondermeer voor schoolsucces, succes in een job of in de loopbaan. Als men bijvoorbeeld iedereen die een score heeft lager dan een kritische grens verder niet in aanmerking neemt, dan neemt men aan dat wie lager scoort dan die kritische grens slecht zou presteren. Het is ook mogelijk dat de testscores zelf of de variabele die ze meten een causale invloed hebben op andere variabelen. Het gaat dan om predictoren met een causale rol.

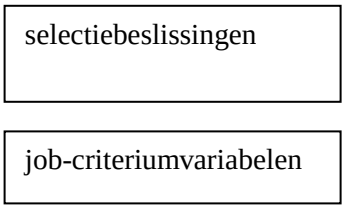
b. Een test is altijd bedoeld om een bepaalde persoonskarakteristiek te meten, verder ook de *bedoelde meetvariabele* genoemd. Die karakteristiek hoeft geen onveranderlijke karakteristiek te zijn, het kan ook gaan om een niveau dat men tijdelijk heeft bereikt of een toestand waarin men tijdelijk vertoeft. Idealiter wordt de testscore geheel bepaald door de bedoelde meetvariabele, maar in de praktijk is het meestal slechts voor een substantieel deel dat de testscore bepaald wordt door de bedoelde meetvariabele. Hoe groter dat deel, des te groter de constructvaliditeit van de test, dat wil zeggen, des te sterker sluit de test aan bij het construct dat men wil meten.



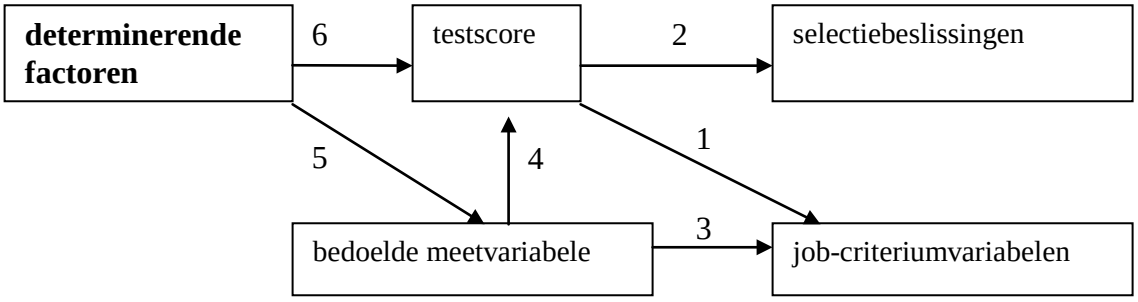
Men kan de band met de bedoelde meetvariabele per item bekijken. In principe speelt de bedoelde meetvariabele een rol in elk item, maar naargelang van het item kan die variabele sterker of minder sterk doorwegen. Het gewicht van de bedoelde meetvariabele

in een item noemt men de “discriminatiegraad” van het item. Hoe groter de *discriminatiegraad* des te sterker differentieert het item tussen hoge en lage waarden van de bedoelde meetvariabele en, des te beter is het item als indicator van de bedoelde variabele. Items hebben naast hun discriminatiewaarde ook nog een moeilijkheidsgraad. Voor juist/fout items is de *moeilijkheidsgraad* het niveau van de bedoelde variabele (de intelligentie) dat nodig is om één kans op twee te hebben om het item juist te beantwoorden.

c. In de context van personeelsselectie zijn de twee belangrijke types van criteriumvariabelen:
(1) *selectiebeslissingen*, zoals de preselectie en de eigenlijke selectie, en (2) gedrag in de job, zoals bijvoorbeeld het prestatieniveau, promoties, het verlaten van de job, enz. , die samen de *job-criteriumvariabelen* worden genoemd. Alleen van wie geselecteerd wordt kan men de waarde op de job-criteriumvariabelen bepalen.



d. Om een volledig beeld te krijgen van de rol die tests kunnen spelen moet ten eerste het onderscheid tussen de testscore en de bedoelde variabele worden ingebouwd in het schema tussen de determinerende factoren en de criteriumvariabelen. Ten tweede moeten de twee types van criteriumvariabelen onderscheiden worden. Op basis daarvan kan men een globaal schema opstellen met de mogelijke invloeden tussen de verschillende bouwstenen.



1.2 Een pragmatische, theoretische en validiteitsbewakende visie op testscores

Een *zuiver pragmatische* aanpak bestaat er in om een test te gebruiken omdat de testscore empirisch predictief is voor de job-criteriumvariabelen waarin men geïnteresseerd is. Men doet dan een beroep op de band die in het schema is aangegeven door pijl 1 die de predictierelatie aangeeft. Pijl 1 geeft de empirische validiteit weer van de test. Er is geen verantwoording van de pijl nodig op grond van een hypothese of theorie. Alleen de feitelijke predictieve waarde van de testscore is van belang. Op grond van de empirische predictierelatie wordt de testscore medebepalend voor de selectiebeslissing, zoals weergegeven door pijl 2. Deze zuiver pragmatische aanpak kan men blind volgen, zonder enige hypothese of theorie. Alleen de pijlen 1 en 2 spelen een rol.

Een *theoretisch geïnspireerde aanpak* bestaat er in om een beroep te doen op hypothesen of een theorie over welke de variabelen zijn die een rol spelen in de job-criteriumvariabelen. De hypothese of theorie betreft de persoonskarakteristieken die bevorderlijk of hinderlijk zijn in de job of de loopbaan. Bijvoorbeeld, bij jobs voor hoger opgeleiden wordt dikwijls aangenomen dat er een minimum aan intelligentie nodig is, naast persoonlijkheidseigenschappen en motivatie. Afhankelijk van de job neemt men aan dat een hogere intelligentie beter is. In een meer gedifferentieerde aanpak bepaalt men ook welke soorten van intelligentie van belang zijn voor de betreffende job of loopbaan. Op basis van dergelijke hypothesen, weergegeven in pijl 3, kiest men bedoelde meetvariabelen en voor deze bedoelde meetvariabelen kiest men tests die constructvaliditeit hebben voor die variabelen, weergegeven in pijl 4. In de theoretisch geïnspireerde aanpak spelen dus ook hypothesen (en theorie) over de job en/of de loopbaan een rol, alsook de constructvaliditeit van tests. Op grond van de pijlen 3 en 4 verwacht men dat de testscore predictief is (pijl 1) en zal de testscore medebepalend zijn voor de selectiebeslissing (pijl 2). Idealiter wordt de predictieve waarde van de testscore ook in de feiten nagegaan, maar dat gebeurt niet altijd. Soms stelt men zich tevreden met de theoretische ondersteuning. Samengevat, spelen in de theoretisch geïnspireerde aanpak de pijlen 1, 2, 3 en 4 een rol.

De basis van dit project is een derde aanpak: de *validiteitsbewakende* aanpak. De aandacht gaat daarbij naar de pijlen 4, 5 en 6. Idealiter verlopen alle invloeden van de determinerende factoren op de testscore via de bedoelde meetvariabele. Dat wil zeggen, als iemand een lagere score haalt op een intelligentietest, dan mag dat alleen maar een reflectie zijn van een lagere intelligentie en niet van iets anders, zoals bijvoorbeeld van de maatschappelijke groep waartoe men behoort, of van de motivatie. Elke invloed op de testscore buiten de bedoelde meetvariabele om is een bedreiging van de constructvaliditeit (pijl 4), want dan spelen naast die bedoelde meetvariabele ook nog andere factoren een rol. Een bedreiging van de validiteit die speciale aandacht vraagt is dat de groep waartoe men behoort een rol speelt in de testscore, los van de bedoelde meetvariabele. Er is dan immers sprake van discriminatie. Pijl 5 geeft de invloed weer van de determinerende factoren op de bedoelde meetvariabele. Pijl 6 geeft de invloed weer van de determinerende factoren op de testscore. De aanwezigheid van pijl 6 is een bedreiging van de validiteit en houdt een discriminatie in als de determinerende variabele

betrekking heeft op de groep waartoe men behoort. Het probleem van DIF en DTF heeft betrekking op pijl 6. De invloed op (de items van) een test kan twee vormen aannemen: een differentiële moeilijkheidsgraad of een differentiële discriminatiegraad. Een differentiële moeilijkheidsgraad betekent dat bepaalde items moeilijker zijn voor de ene groep dan voor de andere. Een differentiële discriminatiegraad betekent dat voor bepaalde items de bedoelde meetvariabele een verschillend gewicht heeft naargelang van de groep. Het is bijvoorbeeld mogelijk dat een item in de ene groep wel een indicator is van intelligentie en in een andere groep niet, of een minder goede indicator. De oorzaak van deze twee vormen van bias (moeilijkheid en discriminatie) kan velerlei zijn: een ander soort voorkennis, een minder goede taalbeheersing, een andere belangstelling. In de validiteitsbewakende aanpak onderzoekt men of naast pijl 4 en 5 niet ook pijl 6 een rol speelt.

1.3 Conceptuele en praktische afbakening van biasonderzoek

Het geschetste kader heeft twee belangrijke implicaties die betrekking hebben of de aflijning van biasonderzoek:

a. Biasonderzoek handelt niet over de invloed die met pijl 5 is weergegeven. Het is mogelijk dat twee bevolkingsgroepen verschillen inzake de meetvariabele zonder dat er van bias sprake is, d.w.z. zonder dat pijl 6 een rol speelt. Als de bedoelde meetvariabele intelligentie is, dan zou dit betekenen dat de ene bevolkingsgroep intelligenter is dan de andere. Vermoedelijk is er een verklaring voor een dergelijk verschil, zoals geringere kansen in het onderwijs, een minder intellectuele opvoeding, genetische aanleg, en dergelijke, maar hoe belangrijk deze invloeden (pijl 5) ook zijn, we rekenen ze niet tot het biasonderzoek (wel tot de differentiële psychologie van de intelligentie). Als men het onderscheid tussen de twee pijlen niet zou maken, dan leidt dat tot onduidelijkheid met als risico dat men niet op de juiste bal speelt als men aan de effecten iets wil doen. Ongewenste effecten die op pijl 6 betrekking hebben kan men oplossen door de tests aan te passen. Ongewenste effecten die op pijl 5 betrekking hebben vergen veel meer, bijvoorbeeld een verandering van het onderwijs.

b. Biasonderzoek kan gebeuren zonder kennis te hebben van selectiebeslissingen of resultaten die geselecteerden behalen in de job of de loopbaan. Het gaat immers alleen maar om de pijlen 4, 5 en 6: het linkse gedeelte uit het schema. Men kan één of meer tests op bias onderzoeken ongeacht wat er verder aan beslissingen op de test volgt en wat de predictieve waarde is van de testscore. Dat een test vrij is van DIF en DTF is een belangrijke verworvenheid die noodzakelijk goede gevolgen heeft voor de selectiepraktijk.

c. We hebben in het geschetste kader geen aandacht gegeven aan de mogelijkheid dat de pijlen in het rechtergedeelte van het schema (1,2 en 3) zelf verschillen naargelang van de groep. Toch kunnen ook dergelijke verschillen voor discriminatie zorgen. Bijvoorbeeld, als een vrouw hogere testscores zou moeten halen dan een man om aangeworven te worden, dan is er een verschil in pijl 2 met discriminatie als gevolg. Dergelijke

praktijken komen voor en zijn afkeurenswaardig, maar we rekenen onderzoek daarover niet tot het biasonderzoek. Het is ook mogelijk dat de bedoelde meetvariabele een ander verband vertoont met prestaties in de job of met het verloop van de loopbaan (een verschil in pijl 3 en dus ook in pijl 1), bijvoorbeeld omdat er verschillende manieren zijn om een job uit te voeren: alternatieve manieren om succes te halen (bijvoorbeeld een mannelijke en een vrouwelijke). Ook dit is een interessant en belangrijk probleem, maar ook dat probleem rekenen we niet tot het biasonderzoek. Geen van beide voorbeelden heeft betrekking op de validiteit van de test.

We hebben niet alleen conceptuele redenen om het onderwerp af te bakenen maar ook twee soorten praktische redenen. De eerste praktische reden heeft betrekking op de remediëring. Als men de geschetste validiteitsbewakende aanpak volgt, kan men zeer doelgericht bepaalde vormen van discriminatie uitschakelen met een grote kans op succes, namelijk die vormen van discriminatie die rechtstreeks betrekking hebben op de tests. De andere vormen van discriminatie vergen een maatschappelijke hervorming die de beperktheid van het project overstijgt. De tweede praktische reden is dat de beschikbaar gestelde middelen een gerichte en afgelijnde benadering vergen om tot concrete tastbare resultaten te komen. Deze keuze betekent geen onderschatting van de andere problemen. Ze is slechts door realisme ingegeven.

1.4 Literatuur over biasonderzoek bij autochtonen versus allochtonen

In Nederland is er dankzij de Commissie Hofstee (1990) de laatste jaren veel onderzoek verricht naar item- en testbias. In Vlaanderen daarentegen heeft deze trend zich niet doorgezet. We bespreken in dit hoofdstuk daarom vooral de resultaten van Nederlands onderzoek.

1.4.1 Literatuur over de verschillen in gemiddelde testscores voor autochtonen en allochtonen

Te Nijenhuis en Van der Flier (1997) onderzochten verschillen in scores op de "General Aptitude Test Battery" (GATB) tussen 806 autochtonen en 1322 allochtonen. Zij constateerden dat allochtonen op alle cognitieve subtests slechter scoorden dan autochtonen (variërend van .40 SD tot 2.07 SD, met een gemiddeld verschil van 1.0 SD). Alle acht tests bleken zowel voor autochtonen als allochtonen dezelfde intelligentiedimensies te meten. Nadat discriminerende items opgespoord en verwijderd waren, bleven de verschillen in gemiddelden tussen de groepen groot.

In een onderzoek van Evers en Lucassen (1991) werden voor 10 van de 12 subtests van de Differentiële Aanleg Test (DAT) significante verschillen in gemiddelde scores tussen autochtone (n=50) en allochtone leerlingen (n=99) gevonden. De onderzoekers doen echter geen uitspraak over de grootte van de gevonden verschillen.

Pieters en Zaal (1991) stelden eveneens verschillen vast tussen allochtonen en autochtonen in een onderzoek naar de Politie Intelligentie Test (PIT). Hierbij concludeerden zij dat allochtonen die in Nederland opleiding hadden genoten beduidend beter presteerden dan allochtonen die hun opleiding in het buitenland hadden gevolgd. Zij voerden ook itembiasonderzoek, maar slechts drie van de 188 items bleken DIF te vertonen.

Van Leest en Bleichrodt (1990) stellen dat hoe meer de culturele achtergrond van kandidaten verschilt met de Nederlands cultuur, hoe lager de testprestaties. Hierbij veronderstellen zij dat de lagere testprestaties voor een deel verklaard kunnen worden door verschillen in Nederlandse taalbeheersing. Hun steekproef bestond uit 207 allochtone kandidaten en er werden 5 verschillende tests gebruikt (Verbaal redeneren, Abstract redeneren, Numeriek vermogen, een test specifiek voor programmeurs en een vocabulairetest).

De Jong en Van Batenburg (1984) onderzochten de intelligentieverschillen bij leerlingen van het zesde leerjaar, gemeten met de verkorte versie van de GALO-test (Groniger Afsluiting Lager Onderwijs) De leerlingen werden opgedeeld naar moedertaal. De verschillen in GALO-scores met de taalgroep Nederlands waren als volgt: Romaans .60 SD lager, Indisch .87 SD lager, Surinaams .93 SD lager en Arabisch 1.13 SD lager.

Resing, Bleichrodt & Drenth (1986) maakten gebruik van de RAKIT (Revisie Amsterdamse Kinder Intelligentie Test) bij autochtone en allochtone kinderen. Zij vonden zowel een verschil bij verbale subtests als bij niet-verbale subtests, zelfs als er gecorrigeerd werd voor sociaal economische status. Zo scoorden Surinaamse, Turkse en Marokkaanse kinderen gemiddeld .55 SD, 1.07 en 1.39 SD lager dan autochtone kinderen. Daarnaast vonden zij ook een significant positief verband tussen verblijfsduur in Nederland en IQscores van allochtone kinderen.

Van den Berg (2001) vond voor alle subtests van de MCT-M (dit is de test die in dit onderzoek ook gebruikt wordt) significante verschillen in scores tussen autochtonen en eerste generatie allochtonen. Tabel 1.1 toont per subtest de gemiddelde score van de autochtonen en van de eerste- en tweede generatie allochtonen. De verschillen in gemiddelden variëren van .20 tot bijna 1 standaarddeviatie in het nadeel van de eerste generatie allochtone groep. De verschillen zijn het grootst bij de verbale tests Woordrelaties en Woordanalogieën en het kleinst bij de subtest Kontroleren. Bij de tweede generatie allochtonen worden vrijwel geen verschillen met de autochtone groep gevonden. Enkel bij de subtest Rekenvaardigheid scoort deze groep significant lager. De gevonden verschillen tussen autochtonen en eerste generatie allochtonen zijn beduidend kleiner dan bij de tests van bovenvermelde onderzoeken. Vermoedelijk wordt het geringere verschil in testscores veroorzaakt door de aanpassingen die bij de ontwikkeling van de MCT-M zijn doorgevoerd om de toepasbaarheid bij allochtone kandidaten te vergroten. Verblijfsduur en de leeftijd waarop men naar Nederland is geëmigreerd hebben beiden een grote invloed op de MCT-M testprestaties van eerste generatie allochtonen. Hoe langer men in Nederland verblijft en hoe jonger men naar Nederland is

geëmigreerd des te hoger de testscores. De leeftijd waarop men naar Nederland is gekomen is daarbij voor de meeste tests meer van belang dan de verblijfsduur.

Tabel 1.1 Overzicht van de gemiddelde subtestscores per groep (autochtonen, allochtonen tweede generatie, allochtonen eerste generatie) in het onderzoek van R.H van den Berg (2001)

Test	Aantal items	Autochtonen (n=857)		Allochtonen tweede generatie (n=135)		Allochtonen eerste generatie (n=648)	
		M	SD	M	SD	M	SD
REKENVAARDIGHEID	30	19.2	7.6	17.8**	6.7	14.5*	6.9
KOMPONENTEN	30	21.0	7.2	21.7	5.2	18.0*	6.3
WOORDRELATIES	45	24.0	9.6	23.9	8.1	15.3*	7.3
CIJFERREEKSEN	30	18.1	5.9	18.3	4.6	15.2*	5.2
KONTROLEREN	100	50.6	16.5	52.4	13.1	47.6*	13.5
SPIEGELBEELDEN	30	13.7	10.4	15.3	10.7	7.9*	8.2
WOORDANALOGIEËN	30	23.9	6.6	24.7	5.7	18.1*	7.4
EXCLUSIE	30	19.6	6.5	20.2	4.4	17.0*	4.7

* significant verschillend van autochtone groep en tweede generatie allochtone groep bij $p<.05$

** significant verschillend van autochtone groep bij $p<.05$

1.4.2 Literatuur over DIF-onderzoek bij autochtonen en allochtonen

De onderzoeken naar DIF bij verschillende etnisch-culturele groepen zijn niet eenduidig. In vrijwel alle onderzoeken wordt DIF geconstateerd, maar het is vaak niet mogelijk om een inhoudelijke verklaring te geven waarom bepaalde items voor de ene groep moeilijker zijn dan voor de andere groep. Wel wordt meestal geconstateerd dat het percentage items dat DIF vertoont hoger is voor verbale tests dan voor niet-verbale tests (Poortinga & Van der Flier, 1988).

Schmitt & Dorans (1990) rapporteren een onderzoek naar DIF bij de Scholastic Aptitude Test (SAT) bij verschillende etnisch-culturele groepen in de VS. De SAT bestaat uit een verbaal en wiskundig gedeelte. Voor drie etnische groepen (Aziaten, Hispanics, Zwarten) werd per deel het gemiddelde percentage items die DIF vertonen gerapporteerd (respectievelijk 16, 15 en 14% voor het verbaal gedeelte en 32%, 2% en 10% voor het Wiskundig gedeelte). De auteurs vermoeden dat het hoge percentage DIF voor de Aziatische groep bij het wiskundige deel gedeeltelijk verklaard kan worden door de verbale inhoud van de items. Een analyse van het verbale gedeelte laat zien dat er een aantal itemkenmerken zijn die de geconstateerde DIF kunnen verklaren. Eerst en vooral is er de inhoud van bepaalde verbale items ("reading comprehension" en "sentence completion"). Items met een inhoud die een bepaalde etnische groep interesseren tonen positieve DIF. Dit betekent dat deze items relatief beter worden beantwoord door deze

groep. Voor analogieën blijkt dat items met woorden die een verschillende betekenis hebben voor de groepen negatieve DIF vertonen. Deze items zijn dus moeilijker voor de allochtone groepen. Analogie-items die woorden bevatten die een associatie oproepen die niet tot de bedoelde analogie behoort, worden door allochtone groepen ook minder goed opgelost.

Een belangrijke vaststelling van Schmitt & Dorans is de 'differential speededness' voor de verschillende etnische groepen. Zwarten en Hispanics met een zelfde SAT verbale totaalscore als blanken, beantwoorden minder items aan het einde van de tests. Hierdoor vertoonden items op het eind van de test meer DIF en was het moeilijk om vast te stellen of dit kwam door de positie van het item of door de inhoud van het item. Ook was het mogelijk dat de "matching variabele", de totaalscore (om bias mee te kunnen vaststellen), door dit verschil in werksnelheid werd beïnvloed.

In Nederland voerde Te Nijenhuis (1997) onderzoek naar DIF in een aantal subtests van de GATB. De analyses worden enkel uitgevoerd op de items die door 90 % van de allochtonen zijn gemaakt. Tabel 1.2 geeft een overzicht van het percentage items dat DIF vertoont per subtest.

Tabel 1.2 Percentages GATB-items die DIF vertonen voor vier verschillende etnisch-culturele groepen in vergelijk met een autochtone groep (naar Te Nijenhuis, 1997)

	Surinamers (n=535)	Antillianen (n=126)	N.-Afrikanen (n=167)	Turken (n=275)
Woordenschat	13%	31%	50%	19%
Driedimensionele Ruimte	20%	13%	25%	13%
Elelementair Rekenwerk	6%	0%	8%	8%

Te Nijenhuis concludeert dat twee van de drie subtests (Woordenschat en Driedimensionele Ruimte) niet vergelijkbaar zijn op itemniveau. Daarnaast stelt hij vast dat de DIF geen grote invloed heeft op de gemiddelde totaalscores. Een probleem bij deze analyses is wel dat er slechts een beperkt aantal items is geanalyseerd, namelijk enkel deze items die door 90% van de groep zijn ingevuld. Dit maakt het moeilijk om een uitspraak te doen over de invloed van DIF in de volledige subtests. Te Nijenhuis vermoedt dat de DIF in de items van Woordenschat en Elementair rekenwerk veroorzaakt wordt door het gebruik van relatief moeilijke woorden die niet tot de woordenschat van de allochtonen behoren. Hij heeft echter geen verklaring voor de DIF in de nonverbale test Driedimensionele ruimte. Net zoals bij Schmitt en Dorans (1990) stelt Te Nijenhuis vast dat de allochtone groepen minder ver komen in de test dan de autochtone groep.

Van den Berg (2001) heeft DIF-onderzoek gedaan op zeven subtests van de MCT-M. In het onderzoek heeft hij twee op itemresponsetheorie gebaseerde DIF-onderzoeksmethoden vergeleken. De resultaten van het onderzoek laten zien dat er bij alle tests sprake is van items die DIF vertonen. Het aantal verschillend functionerende items varieert van 4 (12% van de items) voor de subtest Rekenvaardigheid tot 16 (53%)

voor de subtest Woordanalogieën. De DIF is echter niet duidelijk in het nadeel van één van beide groepen. Sommige items zijn relatief moeilijker voor de ene groep, ander items juist weer voor de andere groep. Kwalitatieve analyses van de items geven geen duidelijke inhoudelijke aanwijzingen voor het verschillend functioneren van de items voor beide groepen. Van den Berg vermoedt dat het verschillend functioneren van de items meer samenhangt met toevallige verschillen in de samenstelling van de onderzoekspopulaties en met de relatieve gevoeligheid van de DIF-methode. Dat de gevoeligheid van de DIF-methode groot is blijkt ook uit het feit dat de effecten van de DIF-items op de totaalscore van elke subtest klein zijn. Bij de verbale tests zijn er veel DIF-items (in 33% van de items van Woordrelaties en 50% van de items van Woordanalogieën). Dit grote aantal lijkt er volgens Van den Berg op te wijzen dat deze tests verschillende dimensies meten.

1.5 Onderzoeksplan

Het onderzoek naar bias bij de doelgroep allochtonen bestaat uit vier fasen. In de eerste fase worden testgegevens verzameld voor relevante intelligentietests. In de tweede fase wordt voor elke test onderzocht welke items een bias vertonen voor de doelgroep in kwestie. Als uit de tweede fase blijkt dat bepaalde items een bias vertonen dan gaan we in de derde fase op zoek naar een inhoudelijke verklaring voor deze bias. Dit kan bijvoorbeeld door de samenhang tussen de bias en bepaalde itemkenmerken te onderzoeken. In een vierde fase, tenslotte, wordt het effect van bias in individuele items op de testscore (aantal juiste antwoorden) en de gecorrigeerde testscore (aantal juiste antwoorden dat gecorrigeerd is voor raden bij meerkeuzevragen) onderzocht. Indien het effect van bias op de testscore substantieel is, dan wordt aangegeven hoe de test eventueel kan worden aangepast. Onze aanpak voor biasonderzoek, zoals die aangewend wordt in de fases 2 tot 4, is gebaseerd op IRT.

De structuur van het rapport kan als volgt worden samengevat: In hoofdstuk 2 beschrijven we de gegevens die verzameld werden in de organisaties
VDAB, SELOR en ABL. In hoofdstuk 3 geven we een theoretisch overzicht van de IRT aanpak voor biasonderzoek. In Hoofstuk 4 bespreken we voor de tests van VDAB het resultaat van verschillende stappen van het biasonderzoek (detectie van DIF, verklaren van DIF, effect van DIF op de testscores). Merk op dat op de testgegevens van SELOR en ABL geen betrouwbaar biasonderzoek kan uitgevoerd worden wegens te kleine steekproeven. In Hoofdstuk 5 onderzoeken we voor de tests van VDAB in welke mate de vaardigheid van respondenten kan verklaard worden als een functie van achtergrondvariabelen zoals leeftijd, opleidingsniveau, enz. In hoofdstuk 6 wordt onderzocht in welke mate gemiddelde testscores van allochtonen en autochtonen verschillen voor de tests van SELOR en ABL. Tot slot geven we in Hoofdstuk 7 een samenvatting van de wetenschappelijke bevindingen van het onderzoek en van de beleidsaanbevelingen die hierbij kunnen geformuleerd worden.

Hoofdstuk 2: Onderzoeksopzet

2.1 Doelgroepafbakening

De bedoeling van dit rapport is te onderzoeken of intelligentietests die veel gebruikt worden voor personeelsselectie discriminerend zijn voor allochtonen versus Vlamingen.
In de literatuur bestaat er weinig consensus over de criteria die gebruikt worden bij doelgroepafbakening wanneer men onderzoek wil doen rond allochtonen. Begrippen als migrant, allochtoon, vreemdeling, minderheid en gastarbeider worden vaak door elkaar gebruikt en er heerst een inhoudelijke begripsverwarring over de gehanteerde concepten. Van de Velde (1992) heeft in de periode 1989-1991 onderzoek gedaan naar de afbakening van de doelgroep bij onderzoek met allochtonen. Hij stelde vast dat in slechts de helft van 81 geregistreerde onderzoeken de doelgroep duidelijk afgebakend werd in termen van nationaliteit en/of herkomst.

Het Koninklijk Commissariaat voor het Migrantenbeleid heeft in 1989 een poging gedaan om doelgroepconcepten te definiëren:

- *Vreemdelingen:* personen die niet de Belgische nationaliteit hebben (juridisch criterium)
- *Allochtonen*: mensen met een andere socioculturele herkomst teruggaand op een ander land van herkomst, ongeacht de huidige nationaliteit (men daalt daarbij meestal af tot de derde generatie, voor zover die mensen zich echter nog als allochtoon willen beschouwen) (sociologisch criterium).
- *Migranten*:
    - mensen die uit een vreemd land komen (demografisch criterium)
    - allochtonen van niet-Europese origine, die meestal - maar niet altijd - in het kader van gastarbeid naar het gastland zijn gekomen en waarbij zich vaak een problematiek van maatschappelijke achterstelling voordoet
- *Etnische minderheden/etnische groepen*: groepen van allochtonen, op één land van herkomst teruggaand, die zich momenteel in het gastland in een achterstands-situatie bevinden.

Het gaat dus vaak om vage begrippen, die in onderzoek moeilijk te operationaliseren zijn. Van Horen en Ramakers (1992) pleiten dan ook voor meer duidelijkheid, vooral rond het begrip 'herkomst'. Zij stellen voor om bij de registratie van de etnisch-culturele positie de volgende gegevens te bevragen: de nationaliteit, de nationaliteit bij de geboorte, het geboorteland, en dit tot de derde generatie terug, langs vaders- en moederszijde. Op basis van deze gegevens kan vervolgens bepaald worden wie men al dan niet tot de allochtone groep laat behoren.
Een kandidaat wordt volgens hen als allochtoon beschouwd *indien hij/zij, minstens één van zijn/haar ouders en/of minstens twee van zijn/haar grootouders bij de geboorte een niet-EU nationaliteit heeft /hebben.* Dit is ook de definitie die door VESOC wordt gehanteerd en die we in dit onderzoek gebruiken.

In dit onderzoek zullen we de gegevens van allochtonen in eerste instantie vergelijken met die van (Vlaamssprekende) autochtonen, verder Vlamingen genoemd. Omdat we meestal over zeer veel gegevens beschikken van Vlaamssprekende autochtonen werd deze groep zeer strikt gedefinieerd als personen met de moedertaal Nederlands die zelf een Belgische nationaliteit hadden bij geboorte, en waarvan beide ouders een Belgische nationaliteit hadden bij geboorte en waarvan de 4 grootouders een Belgische nationaliteit hadden bij geboorte.

Naast *Vlamingen* en *allochtonen* wordt nog een tussengroep gedefinieerd van personen die zelf, ofwel minstens één van de ouders, ofwel minstens twee van de grootouders bij geboorte een niet-Belgische EU nationaliteit hadden. Deze groep wordt verder aangeduid als niet-Vlaamse EU kandidaten.

2.2 Selectie van intelligentietests en onderzoeksopzet

Omdat in selecties de etnische achtergrond van kandidaten niet bevraagd mag worden, zijn er nog geen bestaande datasets voorhanden. De gegevens voor dit onderzoek moesten dus nog verzameld worden. In maart 2003 werd medewerking gezocht met SELOR en ABL voor het biasonderzoek bij de doelgroep vrouwen. Deze twee organisaties waren ook bereid om mee te werken aan het onderzoek naar bias bij de doelgroep allochtonen, op voorwaarde dat de etnische achtergrond op vrijwillige en anonieme basis bevraagd zou worden.

In het biasonderzoek bij de doelgroep vrouwen werden de volgende vier SELOR-tests geselecteerd: ANAVERB, LOGDED, NUMVA en CODES. Deze tests worden vaak afgenomen, ze meten verschillende soorten intelligentie en omdat ze uitvoerig onderzocht werden in het biasonderzoek bij vrouwen, waren ze ook voor dit onderzoek de meest geschikte tests. Wanneer er een grote selectie plaatsvond met één of meerdere van deze tests, dan werd aan de medewerkers van SELOR gevraagd om de kandidaten te motiveren om mee te werken aan het onderzoek. Bij ABL werden alle Nederlandstalige kandidaten die de testen WIMA, TNV en DGEO aflegden, door de consulenten van de verschillende defensiehuizen verzocht deel te nemen aan het onderzoek.
De procedure verloopt in beide instanties als volgt: De kandidaten krijgen de instructie om ná de computertests een vragenlijst in te vullen (zie bijlage 1). Deze vragenlijst bevraagt (1) de huidige nationaliteit, (2) de nationaliteit bij de geboorte en (3) het geboorteland van de persoon in kwestie, van zijn/haar ouders en grootouders. Om een idee te krijgen van de mate waarin allochtonen ingeburgerd zijn, worden nog een aantal bijkomende vragen gesteld, meerbepaald: Hoe lang bent u al in België? Wat is uw moedertaal? Welke taal spreekt u thuis het meest? Met welke bevolkingsgroep voelt u zich het meest verbonden?
De kandidaten waren vrij om al dan niet deel te nemen aan het onderzoek. Om de vragenlijst te kunnen matchen met de scores op de tests, werd gevraagd om de laatste zes cijfers van het rijksregisternummer in te vullen op de vragenlijst. Om de vragenlijst te anonimiseren werd de kandidaat verzocht deze in een enveloppe te steken, dicht te plakken en in de daarvoor voorziene bus te steken.

2.3 Onderzoekspopulatie

Bij SELOR werden er vanaf april 2003 gegevens verzameld. In de vakantieperiode zijn er geen grote selecties geweest, maar in het najaar organiseerde de Vlaamse Gemeenschap enkele grote selecties o.a. voor de functie 'adjunct van de directeur' en 'medewerker' binnen de Vlaamse Gemeenschap.
De dataverzameling voor de doelgroep allochtonen bij ABL werd in juni 2003 gestart in de 5 Vlaamse Defensiehuizen (Hasselt, Antwerpen, Leuven, Brugge en Gent). Vanaf oktober 2003 werden er echter geen nieuwe kandidaten meer getest binnen ABL wegens administratieve en structurele hervormingen. In februari is het onderzoek opnieuw van start gegaan, maar doordat er een tweede (intern) onderzoek binnen ABL liep, kwam ons onderzoek wegens tijdsgebrek op de achtergrond. We hebben dus minder informatie over de kandidaten die in deze periode deelnamen aan de Pin-P.

In tabel 2.1 staan het aantal Vlamingen, het aantal niet-Vlaamse EU kandidaten en het aantal allochtonen die de testen ANAVERB, LOGDED, NUMVA, CODES, WIMA, TNV en DGEO hebben afgelegd én die de vragenlijst hebben ingevuld.

Tabel 2.1: Aantal Vlamingen, niet-Vlaamse EU kandidaten en allochtonen per test

Test	Vlamingen	Niet-Vlaamse EU kandidaten	Allochtonen
LOGDED	2838	122	79
ANAVERB	2968	128	80
NUMVA	1220	31	26
CODES	877	20	16
WIMA	862	77	43
TNV	862	77	43
DGEO	862	77	43

Tabel 2.1 toont dat er weinig allochtone kandidaten deelnamen aan de selecties. Deze groepen zijn te klein om een betrouwbare biasanalyse uit te voeren.
Daarom werd in een tweede stap contact opgenomen met VDAB. Binnen VDAB heeft men in 1999 een onderzoek gedaan naar de kwaliteit van de Multiculturele Capaciteitentest (MCT-M). Deze test bestaat uit 8 verschillende subtests die peilen naar algemeen logisch redeneervermogen, ruimtelijk inzicht, numeriek redeneervermogen, verbaal begripsvermogen en snelheid van redeneren.

In 1999 werd in elk van de dertien lokale klantencentra een collectieve testafname georganiseerd. De participanten waren vooral mensen die in opleiding waren bij de VDAB. Een groot aantal allochtonen werd getest in de opleiding "Nederlands voor anderstaligen". De andere opleidingen waaruit de participanten werden geselecteerd varieerden sterk. De procedure verliep als volgt: Voor het afleggen van de MCT-M werd aan iedereen gevraagd om een (uitgebreid) gegevensformulier in te vullen. Zo verkreeg

men informatie over de nationaliteit (nu en bij geboorte), het geboorteland, de geboortedatum, het geslacht, de verblijfsduur in België, de nationaliteit van de ouders en de grootouders (bij de geboorte), het geboorteland van vader/moeder, de moedertaal, de spreektaal, de gevolgde opleidingen, of men al dan niet een opleiding Nederlands heeft gevolgd en hoe lang, hoeveel jaar betaald werk in België/Buitenland men heeft gehad, de ervaring met intelligentietests/ persoonlijkheidstests. Vervolgens werd nog een korte taaltest afgenomen (speciaal voor de MCT-M ontwikkeld) om te bepalen of de kandidaat voldoende Nederlands kent om de instructies van de MCT-M te begrijpen. Kandidaten die minder dan de helft halen op deze test worden niet uitgenodigd voor het verdere verloop van het onderzoek.
In totaal namen 403 personen deel aan het onderzoek, waarvan 200 Vlamingen en 203 allochtonen. De bedoeling van het VDAB-onderzoek was om na te gaan of de gemiddelde scores van Vlamingen en allochtonen significant van elkaar verschilden, niet of er bias aanwezig was in de items.
De datasets van VDAB werden opgevraagd en er werden een aantal verkennende analyses op uitgevoerd. Hoewel er 403 personen deelnamen aan het onderzoek, zijn er maar van 358 personen volledige gegevens. In deze totale groep zijn er 177 Vlamingen en 161 personen die volgens ons criterium allochtoon zijn. Daarnaast zijn er nog 20 personen, die zelf of één van de ouders, of beide grootouders tot de EU behoren, maar die we niet tot de groep van Vlamingen kunnen rekenen. Binnen de allochtone groep bleek een minderheid hier in België geboren te zijn, de meerderheid is naar België geïmmigreerd na zijn/haar zevende levensjaar.
Om goed biasonderzoek te kunnen doen, moeten we beschikken over een dataset met minstens 200 personen per groep. Daarom werd bovengenoemde procedure in de maanden april, mei en juni 2004 nog eens uitgevoerd. In tabel 2.2 vindt u per subtest het totaal aantal Vlaamse en allochtone kandidaten.

Tabel 2.2 Aantal Vlaamse en allochtone kandidaten per subtest van de MCT-M

Test	Gegevens	
	Aantal Vlamingen	Aantal allochtonen
Rekenvaardigheid	241	289
Komponenten	240	285
Woordrelaties	241	288
Cijferreeksen	242	286
Kontroleren	239	288
Spiegelbeelden	235	285
Woordanalogieën	243	283
Exclusie	242	286

Tabel 2.3 beschrijft de samenstelling van de onderzochte groepen bij VDAB meer in detail. We stellen vast dat er in de totale onderzoeksgroep meer vrouwen aanwezig zijn (66% in de Vlaamse groep en 54% in de groep van allochtonen). De gemiddelde leeftijd is vergelijkbaar in beide groepen (+/- 31 jaar). Binnen de allochtone groep bleek een minderheid (8%) hier in België geboren te zijn, de meerderheid (85%) zijn eerste-generatie allochtonen die naar België geïmmigreerd na hun zevende levensjaar. De

gemiddelde verblijfsduur van deze groep in België is 8 jaar. De Vlamingen zijn daarentegen allemaal in België geboren.
Wanneer we naar het bereikte opleidingsniveau kijken, zien we dat er in de allochtone groep gemiddeld meer hoger geschoolden zijn:41% heeft hoger onderwijs genoten (universitair en niet-universitair) tegenover 15% van de Vlaamse kandidaten.

Verder stellen we vast dat alle Vlamingen Nederlands als moedertaal en spreektaal hebben. Slechts 5% van de allochtonen in onze steekproef heeft Nederlands als moedertaal en 30 % gebruikt Nederlands als spreektaal. Tenslotte blijkt dat Vlamingen meer vertrouwd zijn met het afleggen van tests en het invullen van meerkeuzevragen dan allochtonen. Tabel 2.3 geeft een overzicht van de bovenbeschreven karakteristieken van de onderzoekspopulatie.

Tabel 2.3: Samenstelling onderzoekspopulatie MCT-M

	Vlamingen	Allochtonen
Percentage mannen	34%	46%
Percentage vrouwen	66%	54%
Gemiddelde leeftijd	30 jaar (SD=8.5)	31 jaar (SD=7.6)
Percentage in België geboren	100%	8%
Percentage geïmmigreerd voor zevende levensjaar	0%	7%
Percentage geïmmigreerd nà zevende levensjaar	0%	85%
Gemiddelde verblijfsduur	30 jaar (SD=8.5)	8 jaar (SD=8.5)
Behaald diploma: Basisonderwijs	7%	10%
Lager Secundair Onderwijs	22%	13%
Hoger Secundair Onderwijs	56%	36%
Hoger niet-univ. Onderwijs	11%	14%
Universiteit	4%	27%
Percentage met Nederlands als moedertaal	100%	5%
Percentage met Nederlands als spreektaal	100%	30%
Percentage met testervaring	77%	51%
Percentage met ervaring meerkeuzevragen	87%	75%

Op bovenstaande datasets van SELOR, ABL en VDAB gaan we biasanalyses uitvoeren. Voor de subtests van de MCT-M hebben we waarschijnlijk voldoende kandidaten per groep om betrouwbare conclusies te trekken. Bij de tests van SELOR en ABL hebben we niet voldoende data hebben om DIF betrouwbaar op te sporen. We beperken ons daarom tot het vergelijken van de totaalscores in beide groepen.

2.4 Beschrijving tests

2.4.1 SELOR

2.4.1.1 LOGDED

LOGDED is een test voor transitief redeneren. Hierbij moet men de relatie tussen objecten A en C afleiden op basis van proposities die aangeven hoe A en C gerelateerd zijn aan andere objecten B, D, enz. Er zijn verschillende strategieën mogelijk om tot een conclusie te komen zoals bijvoorbeeld een visuele, een verbale en een algebraïsche aanpak. De test bevat 22 meerkeuzevragen met elk 5 antwoordalternatieven. De kandidaten krijgen 15 minuten om de test op te lossen. Het onderstaand kader toont een voorbeelditem waarbij het juiste antwoord is aangeduid met een pijl.

Voorbeelditem:

A is kleiner dan B; B is kleiner dan C

De relatie tussen A en C is niet te bepalen
→ A is kleiner dan C
A is groter dan C
A is niet groter dan C
A is niet kleiner dan C

2.4.1.2 ANAVERB

Met de test ANAVERB meet men het inductief redeneervermogen aan de hand van verbale analogieën. De kandidaat krijgt telkens twee woordparen aangeboden. Hij dient de relatie tussen deze woordparen te zoeken en aan te vullen. De test bevat 100 meerkeuzevragen met elk 4 antwoordalternatieven. De kandidaten beschikken over 20 minuten om de test op te lossen.

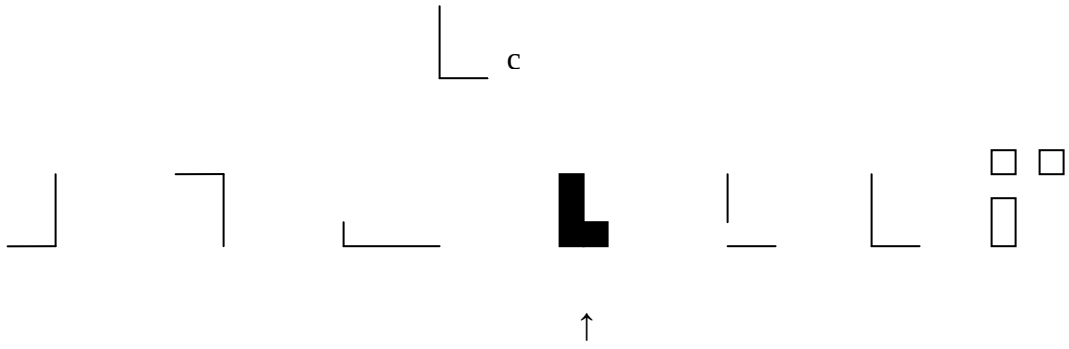
Voorbeelditem:

Lood	pluim	Veder
Zwaar	?	Mooier
		Lucht
		→Licht

2.4.1.3 CODES

Met CODES meet men het vermogen om een code te leren. Elk item bevat één bepaalde code opgebouwd uit letters en leestekens en zeven verschillende figuren. De kandidaat dient de betekenis van de code te achterhalen zodat hij de figuur kan aanduiden die bij de code past. Bij elk item wordt feedback over de juiste oplossing voorzien om het 'leren' van de code mogelijk te maken. De test bevat 74 vragen die dienen opgelost te worden in 45 minuten.
Bij het onderstaande voorbeelditem moeten kandidaten leren dat het een kleine 'c' na een bepaalde letter (in het voorbeeld de letter L) wil zeggen dat deze letter vet wordt weergegeven. Als een code de eerste maal wordt aangeboden moet men raden. Als men geantwoord heeft krijgt men feedback over wat het juiste antwoord was zodat men de code kan leren.

Voorbeelditem:



2.4.1.4 NUMVA

Met NUMVA meet men de vaardigheid om cijferreeksen te vervolledigen, een vorm van inductief redeneren. Elke cijferreeks is opgebouwd volgens een bepaalde logica. De kandidaat dient eerst deze logica te ontdekken en vervolgens uit 4 antwoordalternatieven het alternatief te kiezen dat binnen dezelfde logica de reeks vervolledigt. De test bestaat uit 38 vragen die men moet oplossen in 30 minuten.

Voorbeelditem:

4 6 9 13 18 24 …

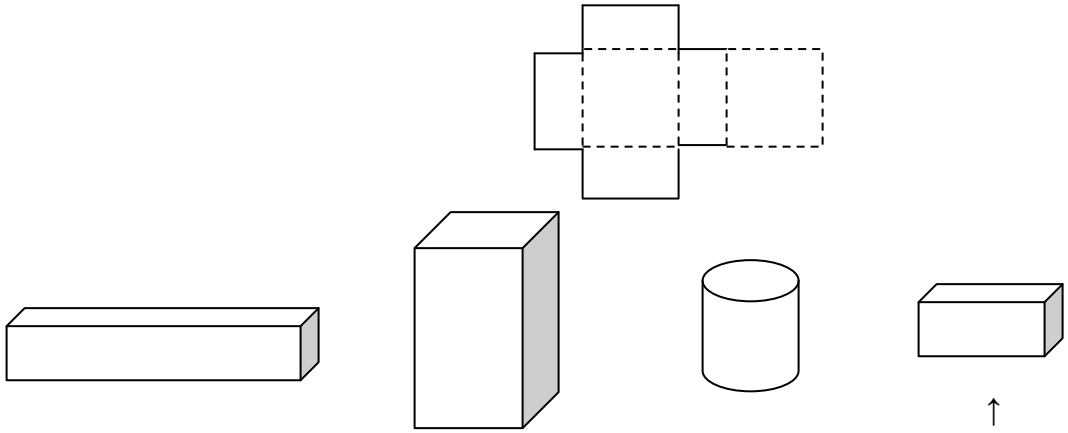
30
6
→ 31
12

2.4.2 ABL

2.4.2.1 DGEO

De DGEO test meet het vermogen tot ruimtelijke visualisatie. De kandidaat krijgt één opengeplooide en vier dichtgeplooide figuren aangeboden. Er wordt gevraagd de figuur aan te duiden die verkregen wordt door de opengeplooide figuur langs de stippellijnen dicht te plooien. De test bevat 40 items die men moet oplossen in 7 minuten.

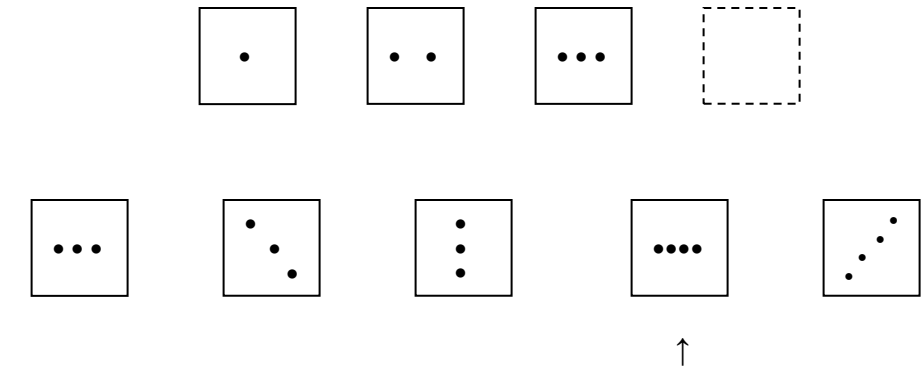
Voorbeelditem:



2.4.2.2 TNV

Met TNV meet men zuivere of vloeiende intelligentie langs niet-verbale weg. Elk item bestaat uit een onvolledige reeks geometrische figuren die een bepaalde samenhang vertonen. De kandidaten moeten de logische samenhang ontdekken en de reeksen aanvullen. Het gaat dus over een taak waarbij inductief redeneren nodig is. De test bevat 50 meerkeuzevragen met elk 5 antwoordalternatieven en moet opgelost worden in 15 minuten.

Voorbeelditem:



2.4.2.3 WIMA

Met WIMA meet men de vaardigheid om numerieke vraagstukken op te lossen De test bevat 23 meerkeuzevragen met telkens 4 antwoordalternatieven. De kandidaten krijgen 30 minuten om de test op te lossen. Men mag hierbij gebruik maken van een kladblad om berekeningen te maken.

Voorbeelditem:

Tijdens een schietoefening wordt aan 9 soldaten het bevel gegeven te schieten in buien van 3 patronen. Hoeveel patronen zullen ze samen opschieten wanneer ze elk 6 buien afvuren.

(A) 27 patronen
(B) 18 patronen
(C) 54 patronen
(D) 162 patronen

2.4.3 VDAB: MCT-M

De MCT-M is een capaciteitentest, die gebruikt wordt om voorspellingen te kunnen doen over de geschiktheid van een individu voor een bepaalde opleiding of functie. Deze test is ontwikkeld en samengesteld om de problemen te verminderen die allochtone kandidaten in Nederland ondervonden bij het oplossen van bestaande capaciteitentests. Bij het ontwikkelen van de MCT-M heeft men aan een aantal aspecten speciale aandacht besteed. Allereerst zijn de instructies uitgebreid en eenvoudig in taalgebruik. Men heeft ook aan elke subtest voorbeeld- en oefenopgaven toegevoegd, zodat de kandidaat op voorhand vertrouwd kan gemaakt worden met de opgaven. Er zijn een aantal niet-verbale subtests opgenomen die vooral uit figuraal materiaal bestaan. Er is de mogelijkheid om een taaltoets af te nemen om na te gaan of de kandidaat voldoende taalkennis heeft om de opgaven te begrijpen. Er zijn allochtone en autochtone normgroepen beschikbaar en men heeft itembiasonderzoek uitgevoerd, zodat de discriminerende items uit de test verwijderd konden worden. De MCT-M bestaat uit 8 subtests met een tijdslimiet. Hieronder worden de verschillende subtests beschreven.

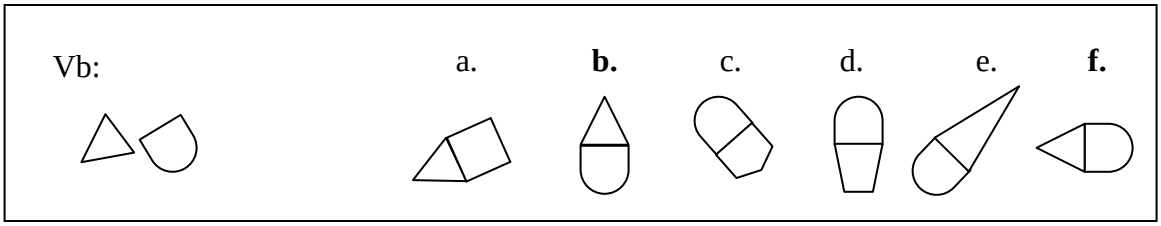
2.4.3.1 Rekenvaardigheid

De rekenvaardigheidstest bestaat uit 30 eenvoudige rekenproblemen met vijf antwoordalternatieven. De rekenproblemen bestaan uit meervoudige en/of gecombineerde optellingen, aftrekkingen, delingen en vermenigvuldigingen. De kandidaten krijgen 5 minuten om de test op te lossen. De test meet enerzijds inzicht in rekenkundige relaties en anderzijds de vaardigheid in het omgaan met getallen (getalbegrip).

Vb: 36 : 6 + 3 = … a.12 b.6 c.7 **d. 9** e.4

2.4.3.2 Komponenten

Deze test bestaat uit 30 items die elk twee kleine figuren bevatten en zes complexe figuren als antwoordalternatief. Twee van deze alternatieven zijn een samenstelling van de twee kleine figuren gemaakt worden. De figuren zijn hierbij aan elkaar gepast en vaak ook gedraaid. Bij deze test gaat het vooral om het mentaal manipuleren en transformeren van figuraal materiaal (spatiale intelligentie). Er is een tijdslimiet van 9 minuten.



2.4.3.3 Woordrelaties

In deze test krijgt de kandidaat 9 minuten om aan te geven welke twee woorden dezelfde of een tegengestelde betekenis hebben. Men meet de mate waarin men de betekenis van woorden kent (woordenschat) en het vermogen om relaties tussen woorden te begrijpen. De 45 items van de test bestaan telkens uit vier woorden.

Vb: a. aardig	**b. goed**	**c. fout**	d.gelijk

2.4.3.4 Cijferreeksen

De test Cijferreeksen meet net zoals NUMVA de vaardigheid om cijferreeksen te vervolledigen. De opeenvolgende getallen van een reeks zijn berekend volgens een bepaalde regel en de kandidaat moet de regel ontdekken en het volgende getalaanvullen. De test bestaat uit 30 reeksen met elk 5 antwoordalternatieven, die men moet op lossen in 15 minuten. De test meet het inductief redeneervermogen om systemen in numeriek materiaal te kunnen ontdekken.

Vb: 3 4 6 9 13 18 ?	a.31	b.23	c.36	d.26	**e.24**

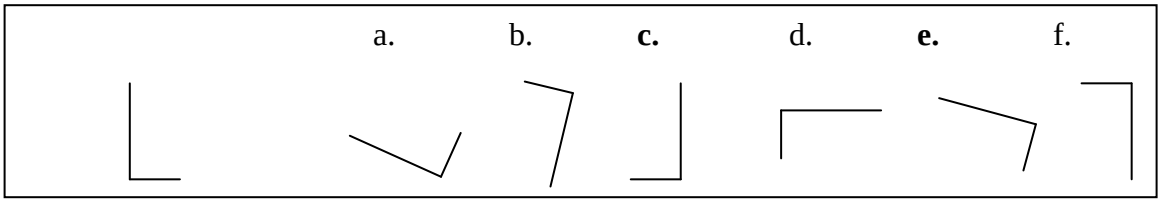
2.4.3.5 Kontroleren

Kontroleren bestaat uit 100 paren van betekenisloze combinaties van letters of getallen. De kandidaat moet aangeven of de combinaties voor en na de streep gelijk of niet gelijk zijn. De test meet perceptuele snelheid en nauwkeurigheid maar ook onder tijdsdruk efficiënt kunnen werken aan een relatief onbekende taak.

Vb:	BPADL – BPADL	**juist**	fout
	848217 – 845217	juist	**fout**

2.4.3.6 Spiegelbeelden

De 30 items van de test Spiegelbeelden bestaan uit een basisfiguur en zes dezelfde maar gedraaide figuren. Twee van de zes alternatieven zijn daarbij ook gespiegeld. Men moet de twee gespiegelde figuren ontdekken. Hiervoor krijgen de kandidaten 15 minuten. De test meet spatiale intelligentie of het vermogen omfiguraal materiaal mentaal te manipuleren en te transformeren.



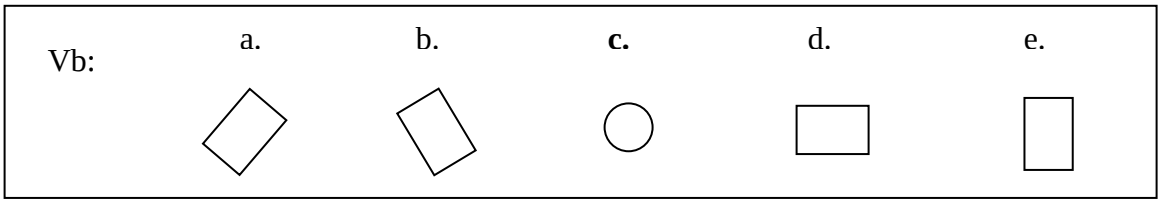
2.4.3.7 Woordanalogieën

De opdracht bij deze test bestaat uit het vinden van twee woordparen die eenzelfde soort relatie hebben (verbale analogieën). De test bestaat uit 30 items met telkens vijf antwoordalternatieven die in 9 minuten moeten opgelost worden. Deze test meet verbaal begrip en verbaal redeneervermogen, dit wil zeggen het kunnen ontdekken van een samenhang of relatie tussen een aantal verbale begrippen.

Vb: Zee – water = land - ?	a. vuur	b. lucht	**c.aarde**	d.wind	e. stad

2.4.3.8 Exclusie

Deze test bevat 30 items met elk vijf figuren. Vier van de vijf figuren hebben een bepaald principe gemeenschappelijk dat men moet ontdekken. De figuur die er niet bij hoort moet aangeduid worden. Hiervoor krijgen de kandidaten 7 minuten. De test meet inductief redeneren.



2.5 Karakteristieken van de datasets

Voor bovenstaande tests hebben we een aantal karakteristieken op een rijtje gezet. Tabel 2.4 geeft per test een overzicht van het aantal voorbeelditems, het aantal items, het aantal antwoordalternatieven, de tijdslimiet, of er al dan niet gecorrigeerd wordt voor raden, het aantal personen per groep en de betrouwbaarheid van de test (coëfficiënt α) (Cronbach, 1951) per groep.

Tabel 2.4 Karakteristieken van de onderzochte datasets

Organisatie	TEST	Aantal vb-items	aantal items	aantal antw-altern.	Tijds-limiet	Gis-cor-rectie	groep	aantal personen	Cronbach's α
SELOR	LOGDED	2	22	5	15 min	ja	Vl	2838	0.76
							EU	122	0.76
							All	79	0.70
SELOR	ANAVERB	2	100	4	20 min	ja	Vl	2968	0.93
							EU	128	0.93
							All	80	0.91
SELOR	CODES	2	74	7	45 min	nee	Vl	877	0.94
							EU	20	0.91
							All	16	0.97
SELOR	NUMVA	2	38	4	30 min	ja	Vl	1220	0.80
							EU	31	0.85
							All	26	0.82
ABL	DGEO	5	40	4	7 min	ja	Vl	862	0.88
							EU	77	0.85
							All	43	0.87
ABL	TNV	5	50	5	15 min	ja	Vl	862	0.88
							EU	77	0.86
							All	43	0.89
ABL	WIMA	2	23	4	30 min	ja	Vl	862	0.88
							EU	77	0.84
							All	43	0.87
VDAB	Rekenvaar-digheid	5	30	5	5 min	nee	Vl	241	0.92
							All	289	0.93
VDAB	Komponten	4	30	6 (2 juist)	9 min	nee	Vl	240	0.89
							All	285	0.91
VDAB	Woord-relaties	5	45	/	9 min	nee	Vl	241	0.89
							All	288	0.88
VDAB	Cijfer-reeksen	5	30	5	15 min	nee	Vl	242	0.86
							All	286	0.85
VDAB	Kontro-leren	5	100	/	4 min	nee	Vl	239	0.97
							All	288	0.97
VDAB	Spiegel-beelden	5	30	6 (2 juist)	15 min	nee	Vl	236	0.96
							All	285	0.96
VDAB	Woordana-logieën	5	30	5	9 min	nee	Vl	243	0.87
							All	283	0.90
VDAB	Exclusie	4	30	/	7 min	nee	Vl	242	0.81
							All	286	0.84

Tabel 2.4 toont dat de meeste tests een hoge tot zeer hoge betrouwbaarheid hebben (α hoger dan .85). Uitzonderingen zijn LOGDED, NUMVA en de Exclusie-subtest van de MCT-M. Zij hebben respectievelijk een betrouwbaarheid van .76, .80 en .81 voor de Vlaamse steekproef en .70, .82 en .84 voor de allochtone steekproef. Verder stellen we vast dat de betrouwbaarheid van de meeste tests ongeveer even hoog is voor allochtonen als voor Vlamingen.

2.6 Verkennende analyses MCT-M

Ter verkenning van de VDAB-gegevens, wordt onderzocht in welke mate het juist kunnen beantwoorden van items in elk van de tests bepaald wordt door enerzijds de tijd die men ter beschikking heeft en anderzijds de moeilijkheid van de items in de veronderstelling dat men tijd genoeg heeft om er aan te werken. Figuren 2.1-2.8 tonen per subtest drie grafieken:

(1) De *globale* proportie allochtonen/Vlamingen die een item juist oplost. Hierbij worden niet ingevulde items fout gerekend.
(2) De proportie allochtonen/Vlamingen die het item hebben ingevuld. Deze figuur geeft informatie over een mogelijk verschil in snelheid tussen beide groepen.
(3) De proportie allochtonen/Vlamingen die het item juist hebben opgelost gegeven dat ze het hebben ingevuld. Deze figuur geeft aan of een item moeilijker is voor beide groepen in de veronderstelling dat men de tijd genomen heeft om het op te lossen en kan dus wijzen op een onderliggend verschil in vaardigheid.

Uit de figuren blijkt dat snelheid en vaardigheid in verschillende mate de globale proportie juist bepalen (in de veronderstelling dat niet ingevulde items fout worden gerekend). Aan het ene einde van het continuüm bevindt zich de test “rekenvaardigheid”: Bij deze test is er wel een snelheidsverschil tussen beide groepen, maar weinig verschil in vaardigheid. Personen in beide groepen maken ongeveer alle ingevulde items juist maar daarnaast geldt dat naarmate de test vordert allochtonen minder items invullen omdat ze trager werken. De verschillen in de globale proportie juist tussen beide groepen wordt dus voornamelijk bepaald door de tijdslimiet en door het verschil in snelheid tussen beide groepen. Aan het andere einde van het continuüm bevindt zich de test spiegelbeelden. Bij deze test hebben maken allochtonen meer fouten bij de items die ze invullen maar is het verschil in snelheid niet zo groot.

Om op verder te onderzoeken in welke mate de globale proportie juist afhangt van snelheid en vaardigheid wordt een lineaire regressie uitgevoerd met als onafhankelijke variabele het verschil in globale proportie juist voor Vlamingen en allochtonen en als predictoren (1) het verschil in proportie ingevuld voor beide groepen en (2) het verschil tussen beide groepen in proportie juist van items die ingevuld werden. De resultaten van deze analyses worden weergegeven in Tabel 2.5

Tabel 2.5 toont dat het verschil in globale proportie juist tussen Vlamingen en allochtonen verklaard wordt door zowel een verschil in vaardigheid, als een verschil in

snelheid. Vlamingen zijn in het algemeen vaardiger dan allochtonen en zij werken sneller. Toch is er een verschil in bijdrage van beide variabelen, afhankelijk van de subtest. Sommige subtests meten vooral een verschil in snelheid (Rekenvaardigheid, Kontroleren, Komponenten). Voor deze tests zijn alle items ongeveer even moeilijk en is het dus vooral een kwestie van snel antwoorden. Andere tests meten voornamelijk een verschil in vaardigheid (Spiegelbeelden, Woordrelaties en Woordanalogieën). Bij deze tests met relatief moeilijke items is het vooral belangrijk dat men over de vaardigheid beschikt om deze items op te lossen. Tot slot zijn er tests waar zowel verschillen in snelheid als verschillen in vaardigheid een rol spelen (vb, exclusie, cijferreeksen).Vlamingen zijn in het algemeen vaardiger dan allochtonen en zij werken sneller, waardoor zij meer kans hebben om items juist op te lossen. Men zou kunnen overwegen om allochtonen meer tijd geven voor het oplossen van de tests. Maar als dit betekent dat ongeveer iedereen alle items juist oplost, dan kan men wel geen onderscheid meer maken tussen vaardige en minder vaardige personen.

Tabel 2.5 Bijdrage van verschil in vaardigheid en verschil in snelheid aan verschil in proportie juist bij Vlamingen en allochtonen.

Test	Verschil in vaardigheid		Verschil in snelheid		R^2
	gewicht	p-waarde	gewicht	p-waarde	
Rekenvaardigheid	0,10	0,0044	0,99	<,0001	0,97
Kontroleren	0,14	<,0001	0,96	<,0001	0,96
Komponenten	0,30	0,0003	0,79	<,0001	0,87
Exclusie	0,61	<,0001	0,46	<,0001	0,91
Cijferreeksen	0,56	0,0003	0,38	0,0078	0,76
Woordanalogieën	0,79	<,0001	0,31	<,0001	0,98
Spiegelbeelden	0,84	<,0001	-0,39	0,0001	0,79
Woordrelaties	0,95	<,0001	0,02	0,7214	0,91

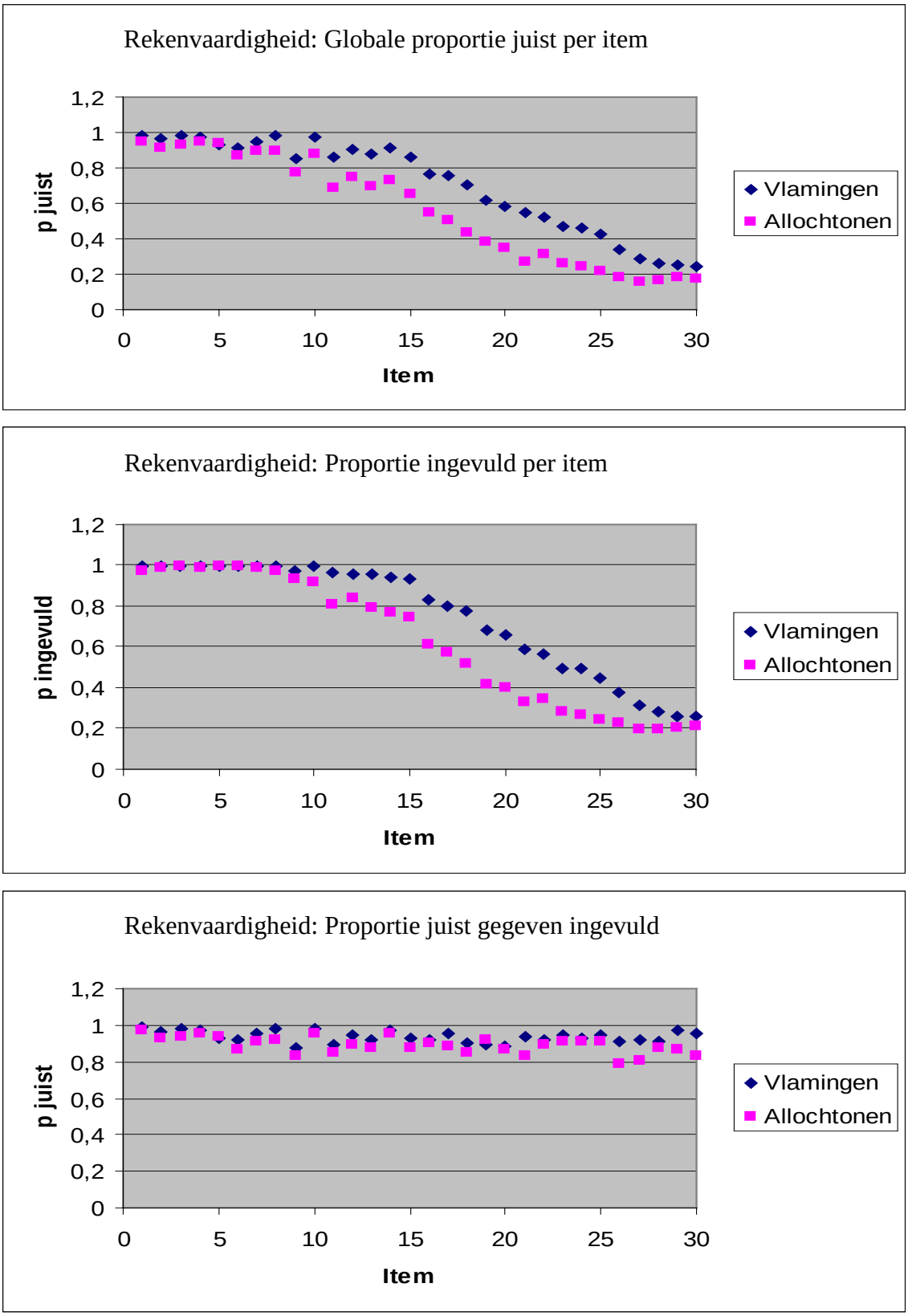
2.7 Scoringsvoorschriften

De bovenvermelde verkennende analyses tonen dat er een verschil in werksnelheid is tussen allochtonen en Vlamingen. Om verder te onderzoeken in welke mate snelheid en vaardigheid een rol spelen bij de DIF resultaten hebben we de DIF analyses voor alle subtests uitgevoerd op datasets met twee soorten coderingen: (1) een dataset waarbij de items die niet worden ingevuld fout worden gerekend en (2) een dataset waarbij de items die niet worden ingevuld worden beschouwd als ontbrekend. Uit een vergelijking van beide analyses blijkt dat het hoofdeffect en de DIF parameters over het algemeen weinig verschillen. We zullen daarom alleen het resultaat rapporteren van de analyses waarbij niet-ingevulde items worden fout gerekend. Een uitzondering waarbij de twee soorten analyses wel verschillende resultaten opleveren is de subtest Kontroleren. Dit wordt verder in Hoofdstuk 4 besproken.

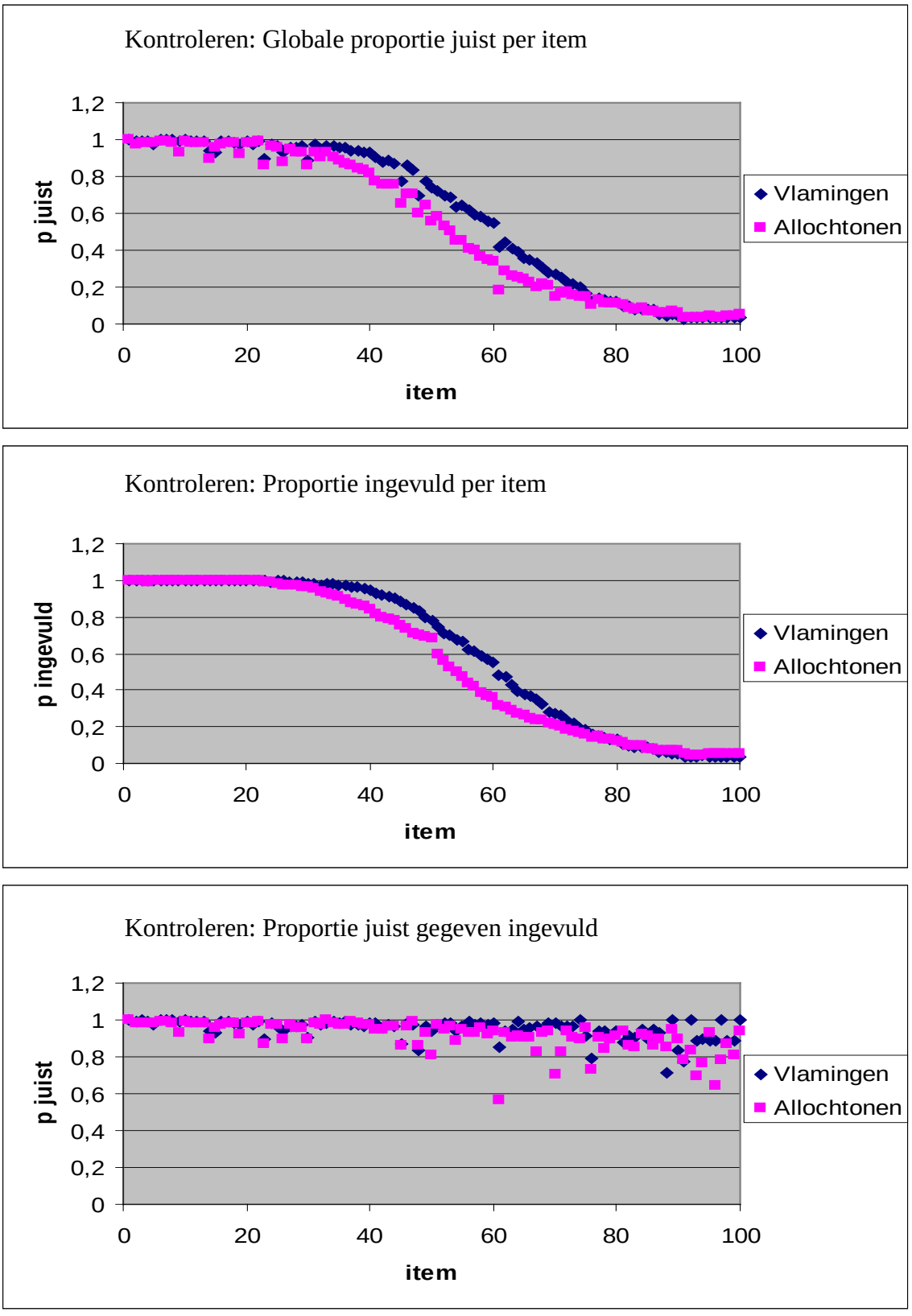
Bij SELOR en ABL beschrijft men (behalve bij CODES) de prestatie van een kandidaat op de test aan de hand van een totaalscore die corrigeert voor het feit dat men meerkeuzevragen ook juist kan oplossen door te raden. Deze gecorrigeerde totaalscore wordt berekend op basis van het aantal juiste, het aantal foute en het aantal onbeantwoorde items:

Gecorrigeerde totaalscore = aantal juiste antwoorden – (aantal foute antwoorden /(aantal antwoordalternatieven -1))

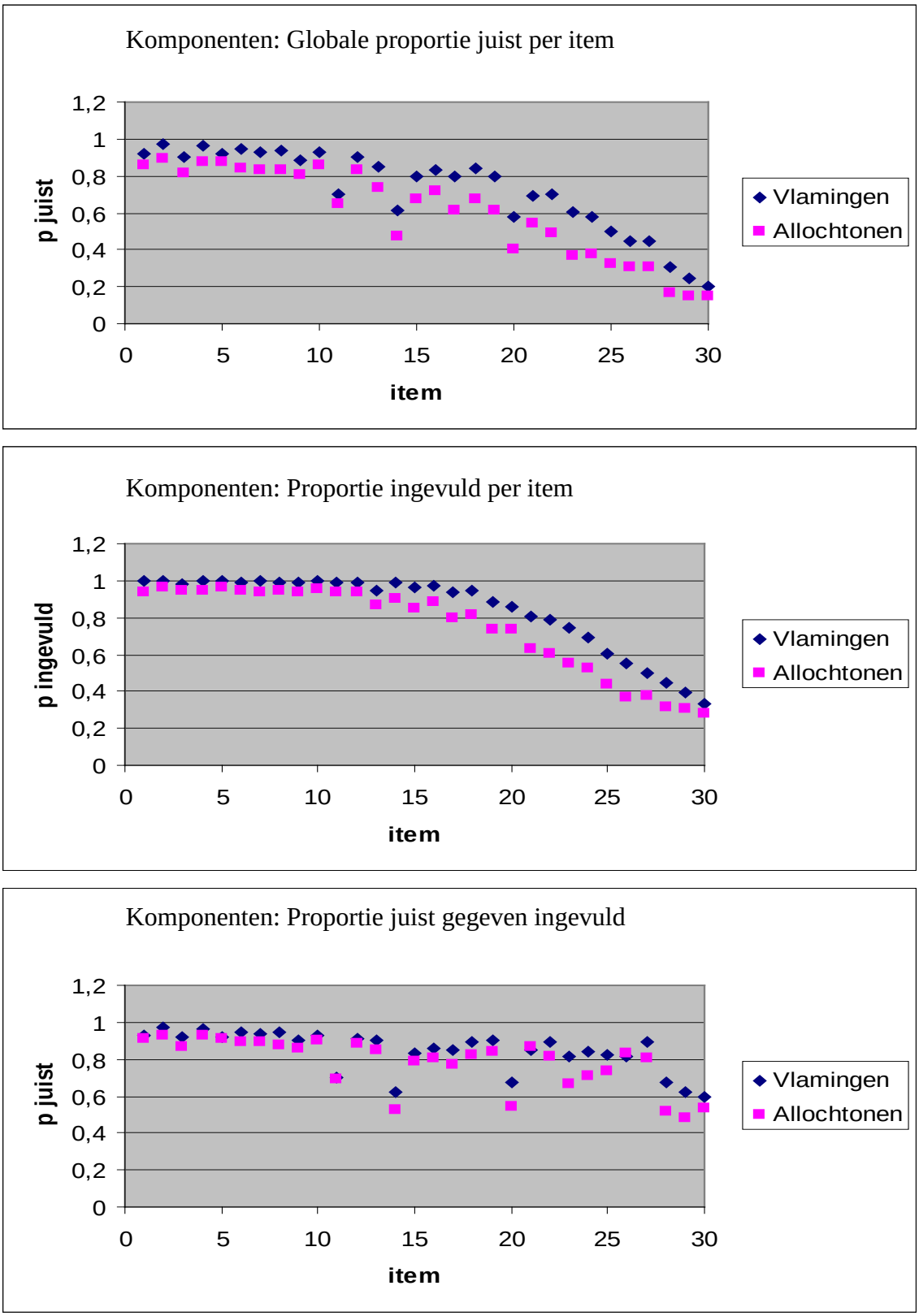
Bij de subtesten van de MCT-M wordt er niet gecorrigeerd voor raden. Wel wordt raden soms bemoeilijkt doordat er twee juiste antwoorden zijn (vb. Spiegelbeelden, Komponenten). Men moet beide antwoorden correct hebben om het item juist op te lossen.



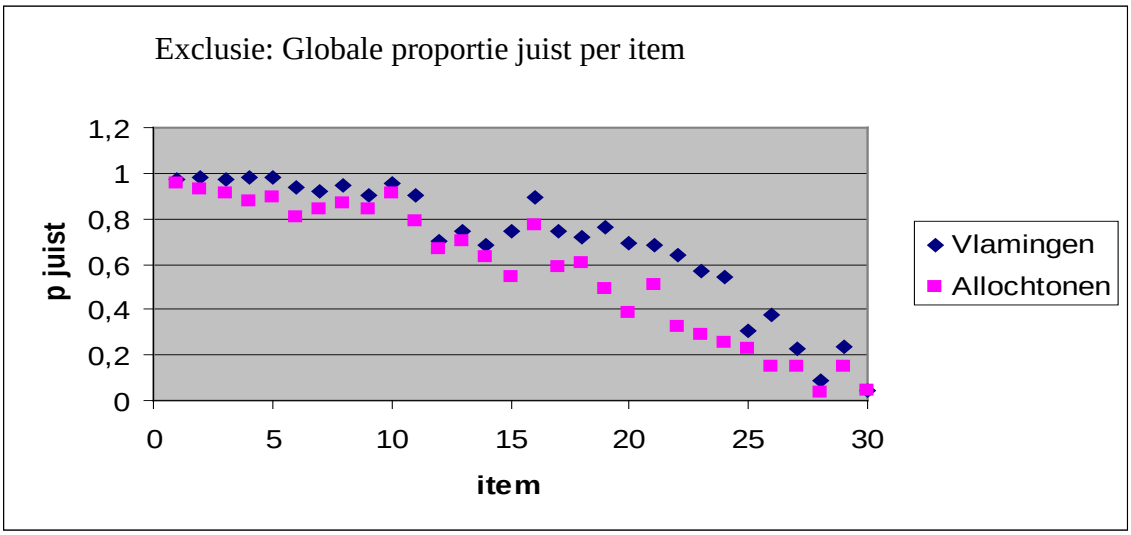
Figuur 2.1 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Rekenvaardigheid

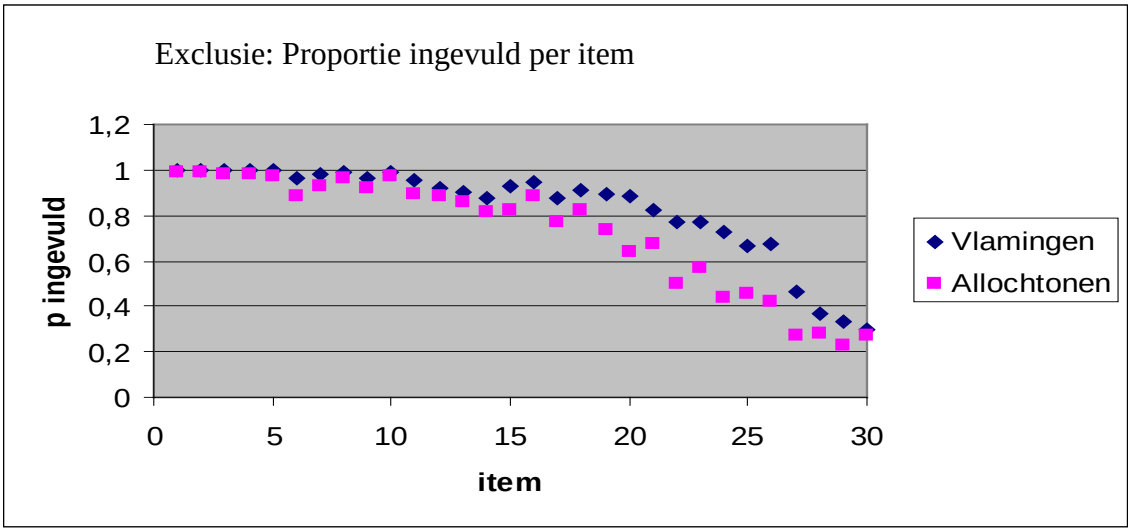


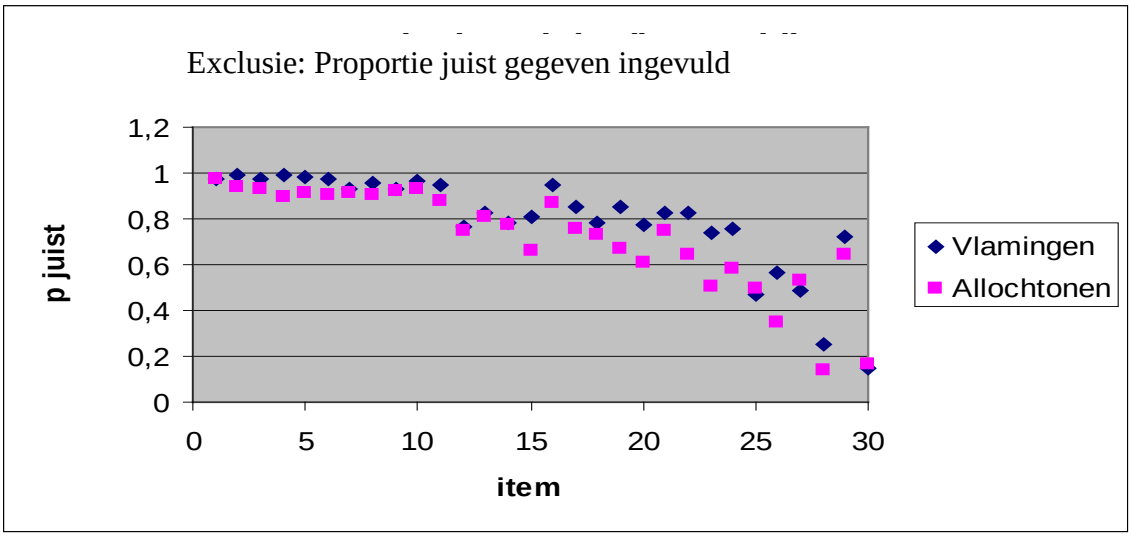
Figuur 2.2 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Kontroleren



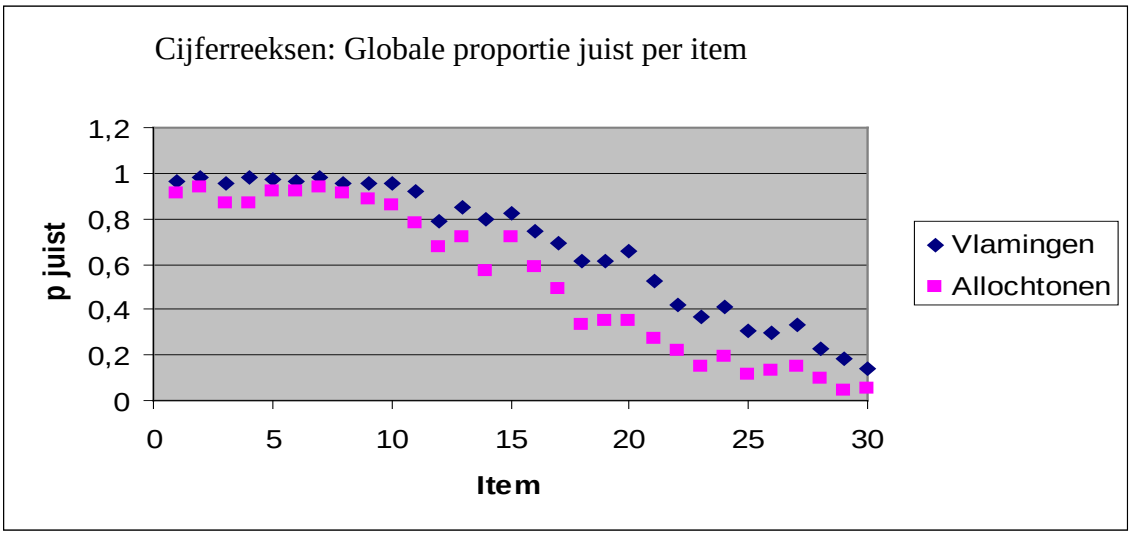
Figuur 2.3 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Komponenten

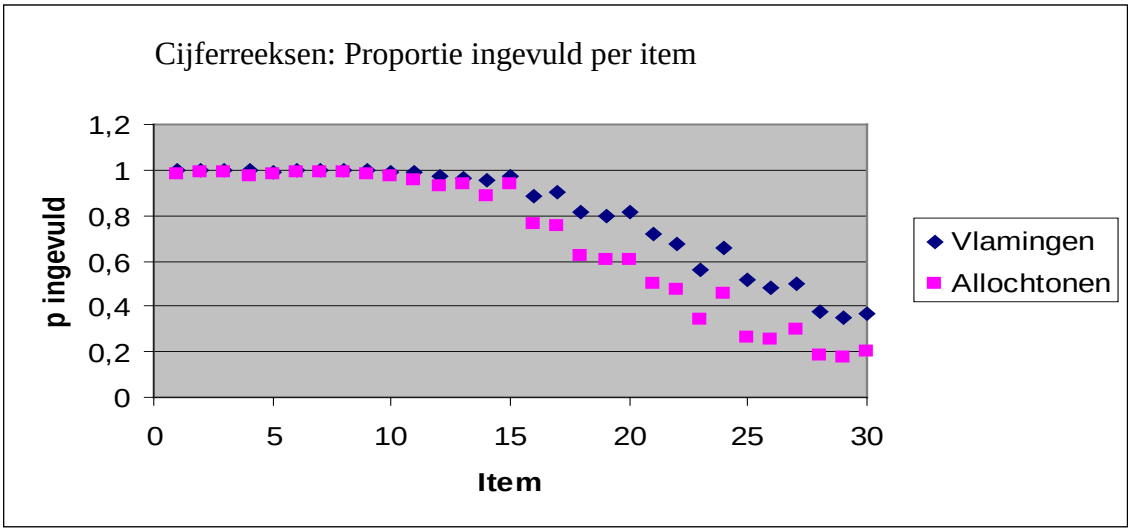


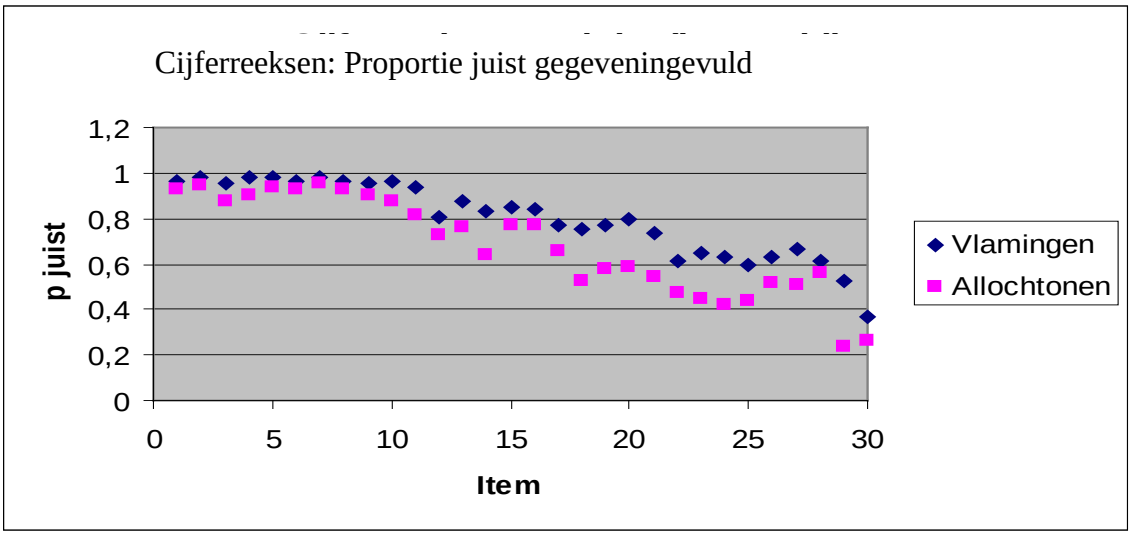




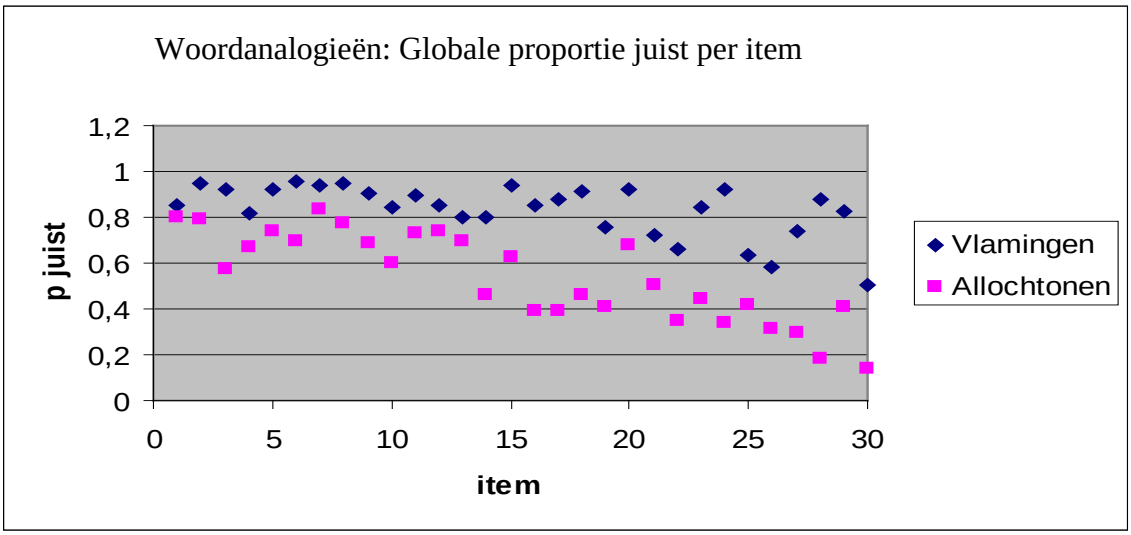
Figuur 2.4 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Exclusie

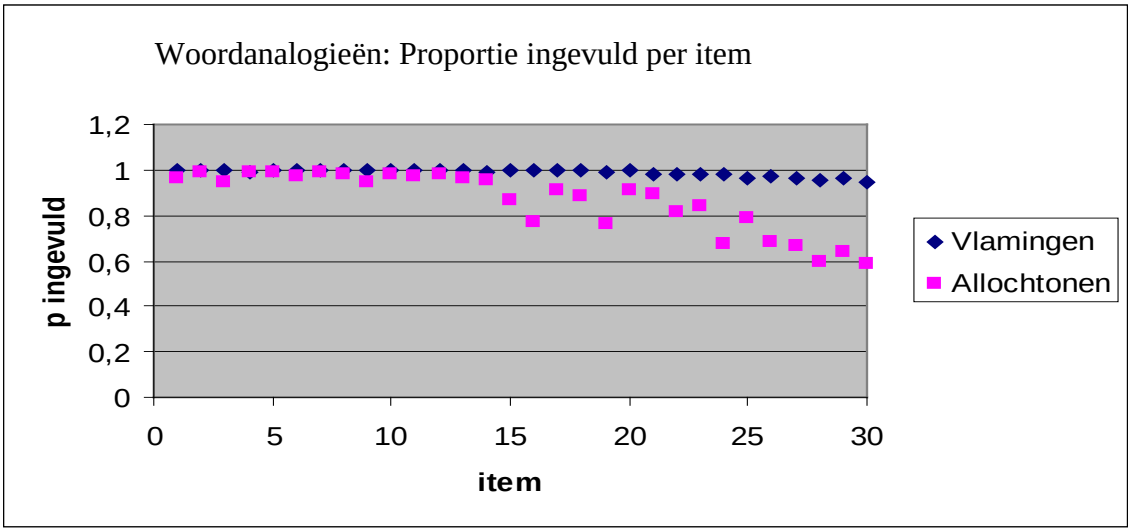


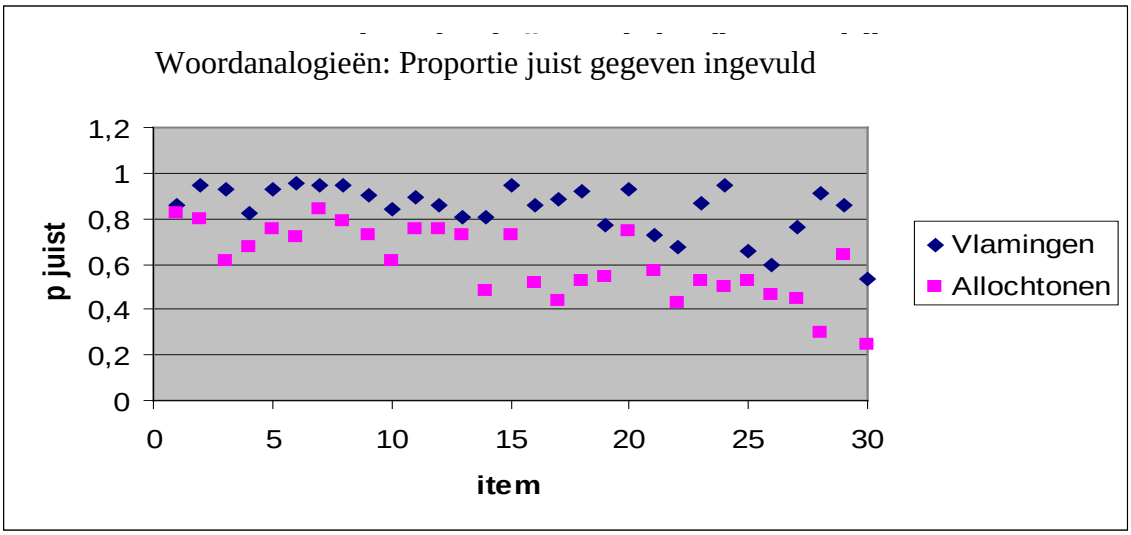




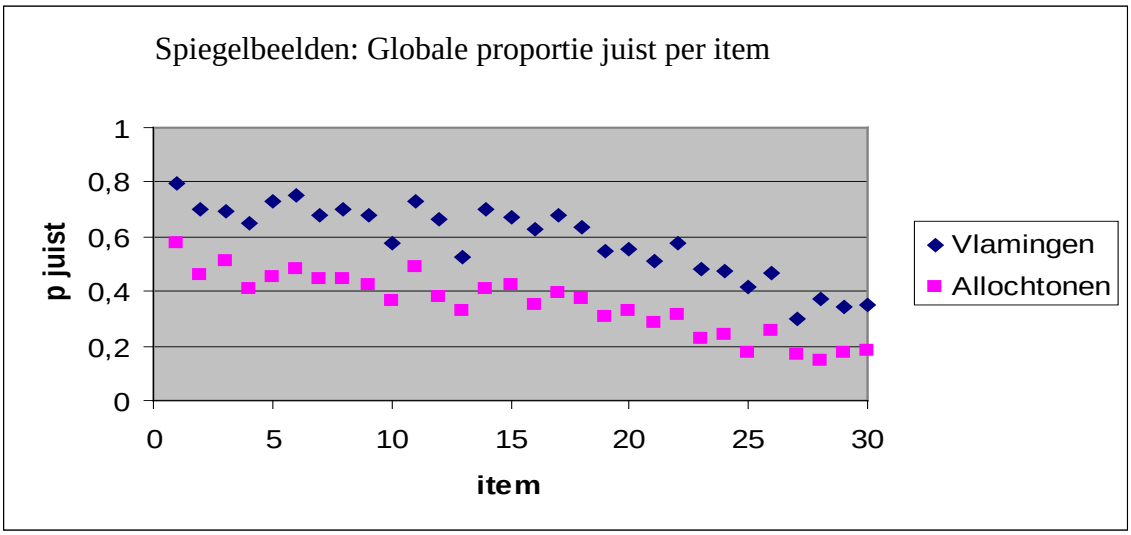
Figuur 2.5 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Cijferreeksen

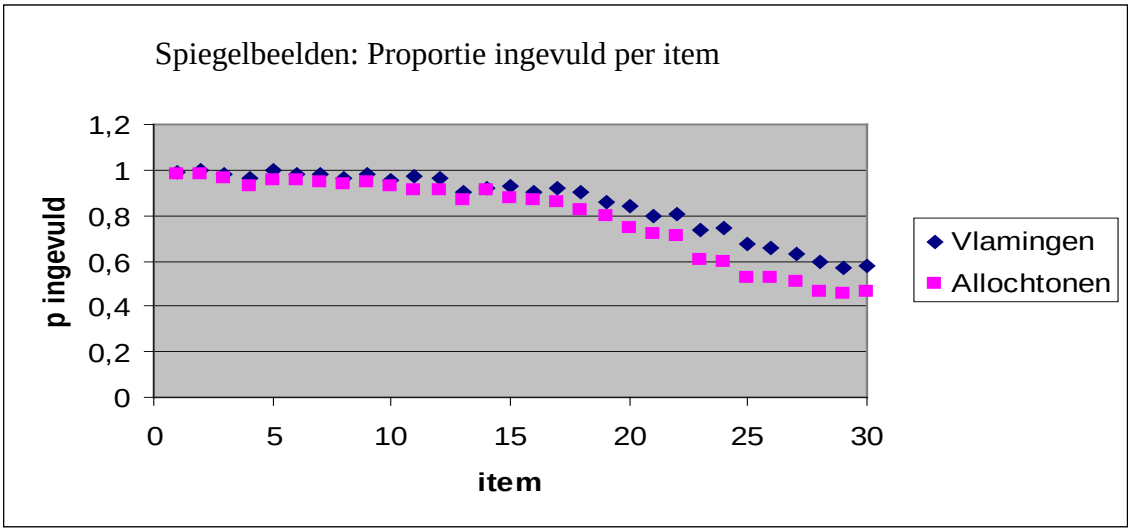


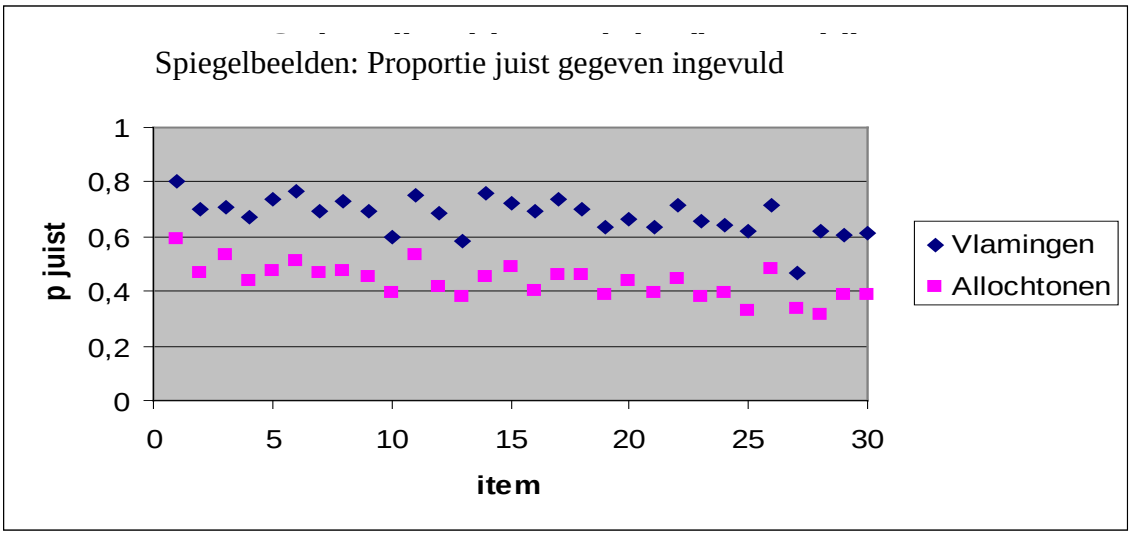




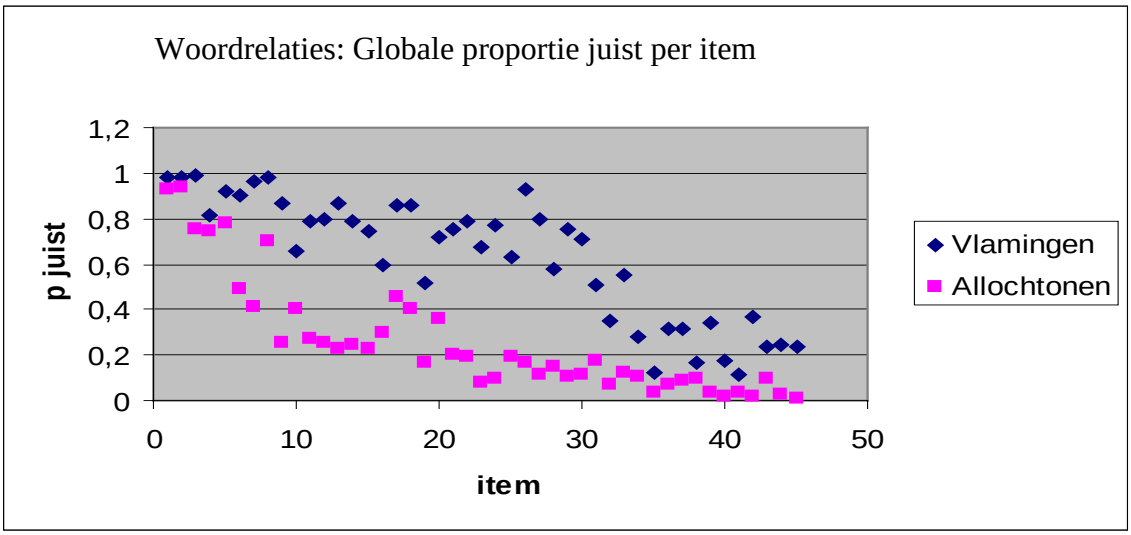
Figuur 2.6 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Woordanalogieën

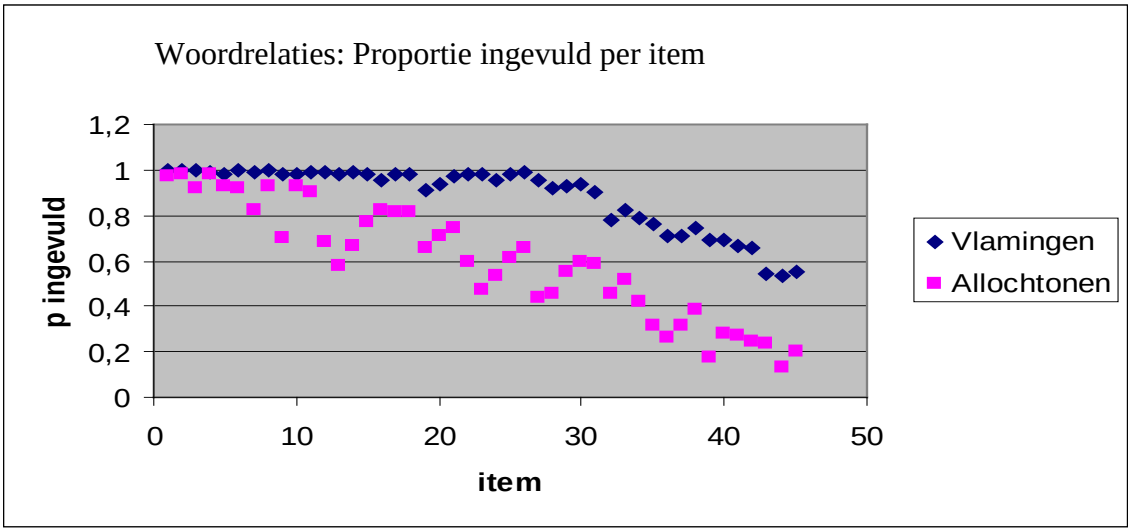


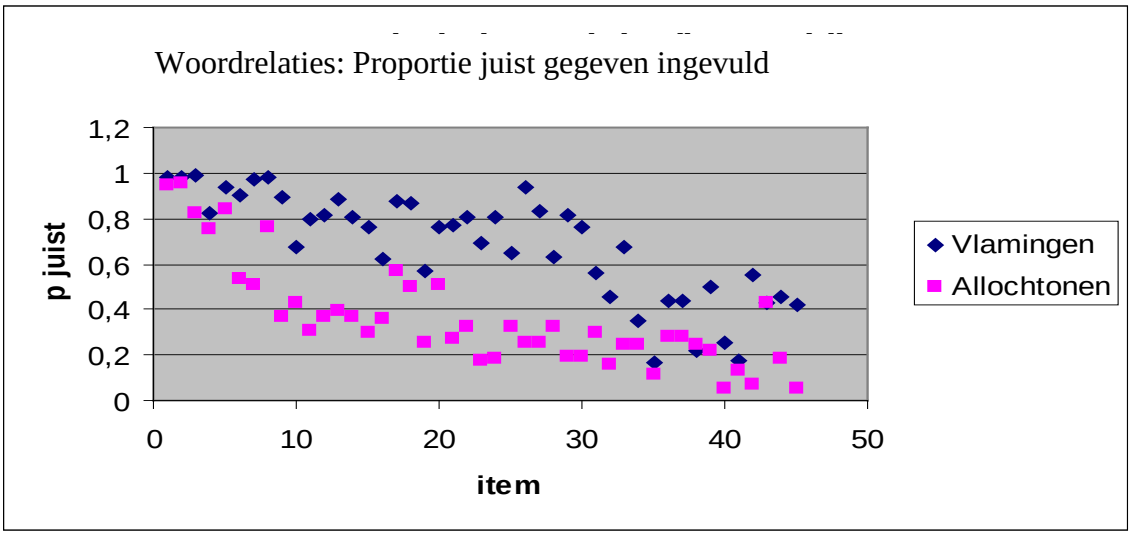




Figuur 2.7 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Spiegelbeelden







Figuur 2.8 Overzicht van verschil in snelheid, verschil in vaardigheid en verschil in proportie juist bij Woordrelaties

Hoofdstuk 3: Een IRT Benadering voor biasonderzoek

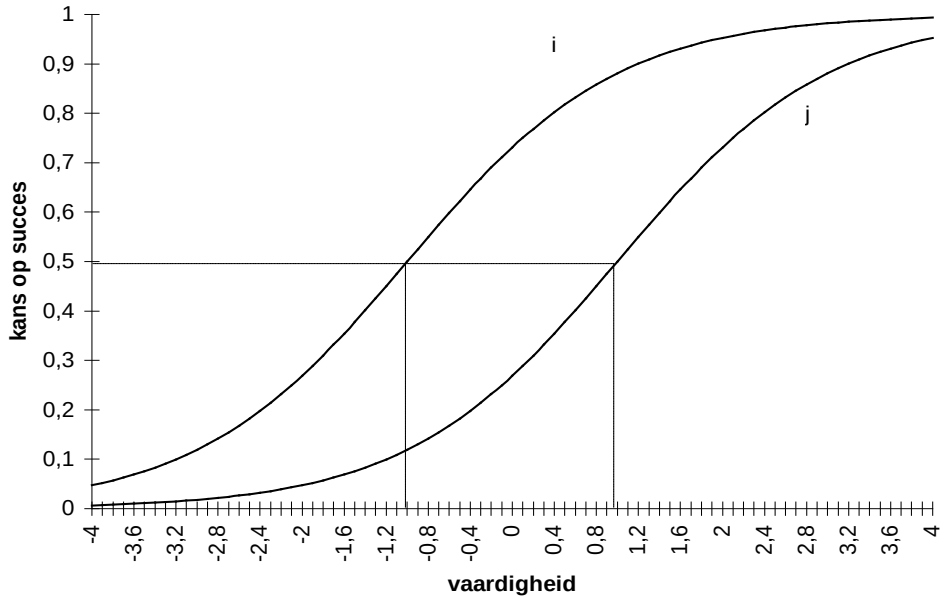
3.1 Itemresponsmodellen

Voor de analyse van testgegevens en meer specifiek voor biasonderzoek kan gebruik gemaakt worden van modellen die ontwikkeld zijn binnen de itemresponstheorie. Itemresponsmodellen beschrijven op basis van de geobserveerde testgegevens de kans dat een persoon met een bepaalde vaardigheid een bepaald item juist oplost. Meerbepaald wordt deze kans gemodelleerd als een functie van de vaardigheid van de persoon en van bepaalde itemkenmerken zoals bijvoorbeeld de moeilijkheidsgraad van het item. Een zeer algemeen model dat veel gebruikt wordt voor de analyse van testgegevens is het 3-parameter logistisch (3PL) model. Duiden we het antwoord van persoon p op item i aan met de binaire variabele Y_{pi} (Y_{pi}=1 als persoon p item i juist beantwoordt en 0 als dat niet zo is), dan stelt het 3PL dat persoon p item i juist beantwoordt met kans:

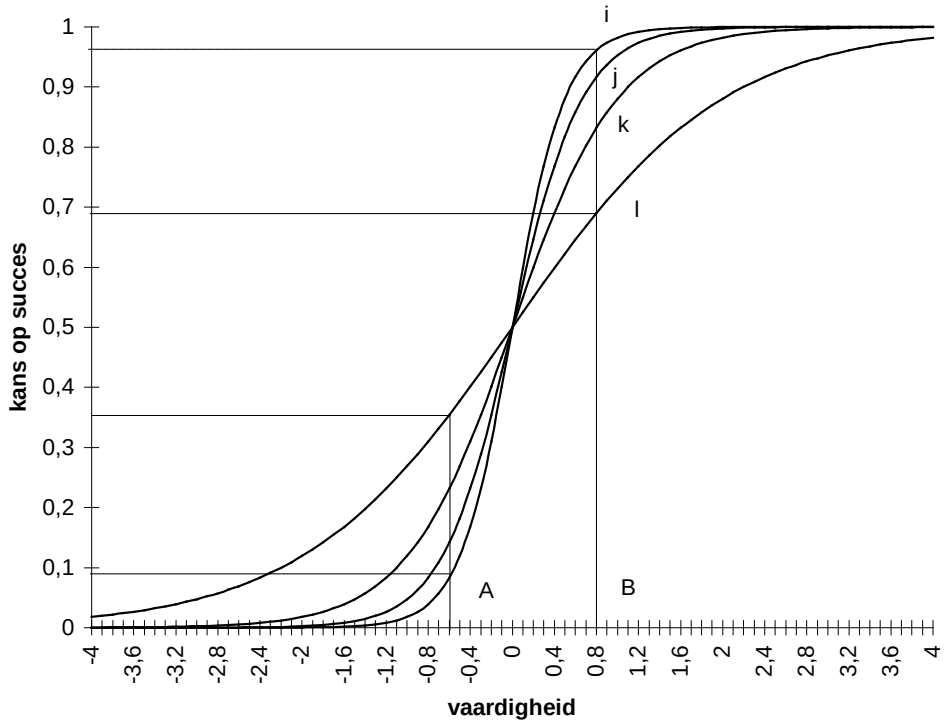
$$\Pr(Y_{pi}=1 \mid \theta_p, \alpha_i, \beta_i, \gamma_i) = \gamma_i + (1-\gamma_i)\frac{\exp\left[\alpha_i(\theta_p-\beta_i)\right]}{1+\exp\left[\alpha_i(\theta_p-\beta_i)\right]} \qquad \textbf{(3.1)}$$

Hierbij verwijst θp naar de positie van persoon p op het latente vaardigheidscontinuum θ. Het verband tussen de vaardigheid van een persoon en de succeskans kan grafisch worden voorgesteld als een itemresponsfunctie (IRF) (zie Figuren 3.1, 3.2 en 3.3). Uit de figuren blijkt dat de succeskans een stijgend S-vormige functie is van de vaardigheid. Met andere woorden personen met een hoge score op θ (gesitueerd aan de rechterkant van de schaal) hebben meer kans om een item juist op te lossen dan personen met een lage score (gesitueerd aan de linkerkant van de schaal). De itemparameters αi, βi en γi hebben een specifieke interpretatie.

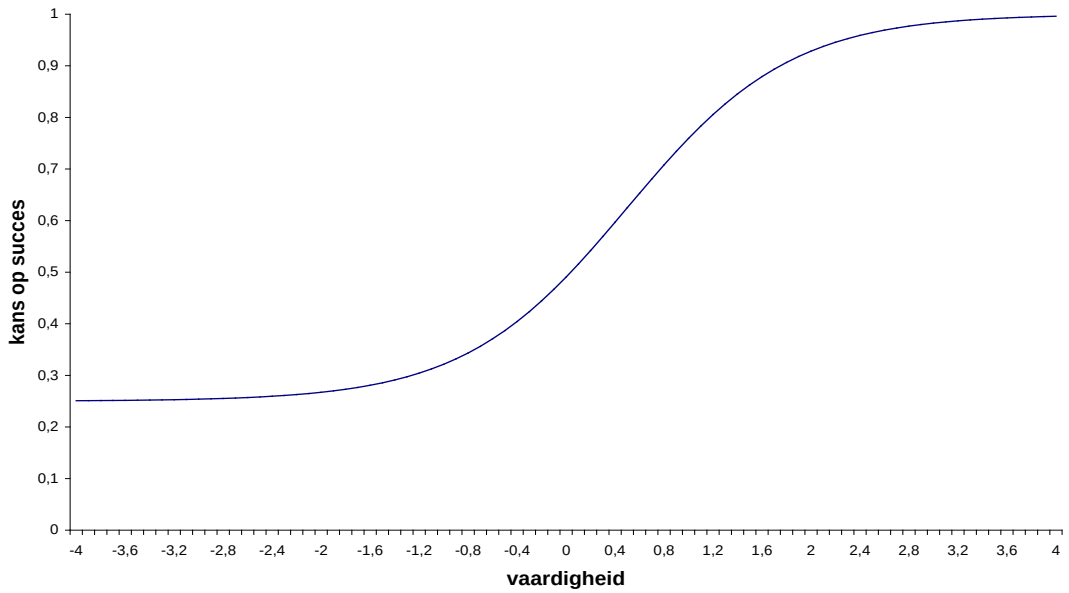
De parameter γi, ook wel raadparameter genoemd, is de kans om het item juist op te lossen als men een oneindig lage vaardigheid heeft. In de grafische voorstelling is deze parameter de linkerasymptoot van de itemresponsfunctie. Bij open vragen is het redelijk om deze parameter op voorhand gelijk te stellen aan 0 (zoals bijvoorbeeld het geval in de Figuren 3.1 en 3.2). Bij meerkeuzevragen waar men op toeval het juiste alternatief kan kiezen is het echter aangewezen om de parameter te schatten op basis van de gegevens of om de parameter gelijk te stellen aan één gedeeld door het aantal antwoordalternatieven (de kans om op toeval juist te antwoorden). Figuur 3.3 toont een itemresponsfunctie voor een meerkeuzevraag met 4 antwoordalternatieven en raadparameter gelijk aan 0.25.
De parameter βi kan geïnterpreteerd worden als de moeilijkheidsgraad van het item. Als de raadparameter gelijk is aan 0 dan geldt dat personen met dezelfde positie op de schaal als het item (en dus θ=β) een kans van 0.5 hebben om het item juist op te lossen. Meer in het algemeen geldt bij het 3PL dat voor θ=β de kans op succes gelijk is aan γ+(1-γ)*0.5. Naarmate items moeilijker worden bevinden ze zich meer naar rechts op de schaal (zie Figuur 3.1).



Figuur 3.1 Itemresponsfuncties voor items met gelijke discriminatiegraad (α i = αj=1) en verschillende moeilijkheidsgraad βi = -1 en βj = 1.



Figuur 3.2 Itemresponsfuncties voor items met gelijke moeilijkheidsgraad βi = βj = βk = βl = 0 en verschillende discriminatiegraad αi = 4, αj = 3, αk = 2 en αl = 1.



Figuur 3.3 Itemresponsfunctie voor een meerkeuzevraag met 4 antwoordalternatieven en een raadparameter gelijk aan 0.25.

De parameter α, ook wel discriminatiegraad genoemd, beschrijft de sterkte van het verband tussen de latente trek θ en de succeskans voor het item. Naarmate α groter wordt stijgt de succeskans sneller in functie van θ (zie Figuur 3.2). Met andere woorden, het item kan beter een onderscheid maken tussen personen met een hoge en een lage vaardigheid. We merken nog op dat het deel binnen de exponent in formule (3.1) soms geherformuleerd wordt als $\alpha i(\theta p-\beta i)= \alpha i\theta p-\delta i$ waarbij δi de itemdrempel genoemd wordt.

Op basis van de geobserveerde gegevens is het mogelijk om met bepaalde statistische procedures, zoals bijvoorbeeld de NLMIXED procedure van SAS, parameterwaarden te bepalen die voor het gekozen IRT model optimaal de gegevens beschrijven. Hierbij maakt men meestal de veronderstelling dat de persoonsparameters (θ) een bepaalde verdeling volgen, bijvoorbeeld, een normale verdeling.

3.2 Differential item functioning

In de context van itemresponstheorie spreken we van *differential item functioning (DIF)* als de itemresponsfunctie een verschillend verloop kent in verschillende groepen (Lord, 1980). Anders gezegd, de afwezigheid van DIF impliceert dat voor elk punt op het vaardigheidscontinuum de succeskansen voor beide groepen gelijk zijn. Als Z de groepsvariabele aanduidt (meerbepaald Z=0 voor Vlamingen, Z=1 voor allochtonen), kan dit formeel worden weergegeven als:

$Pr(Y_{pi}=1|\theta,Z=0) = Pr(Y_{pi}=1|\theta,Z=1)$ voor alle waarden van θ.

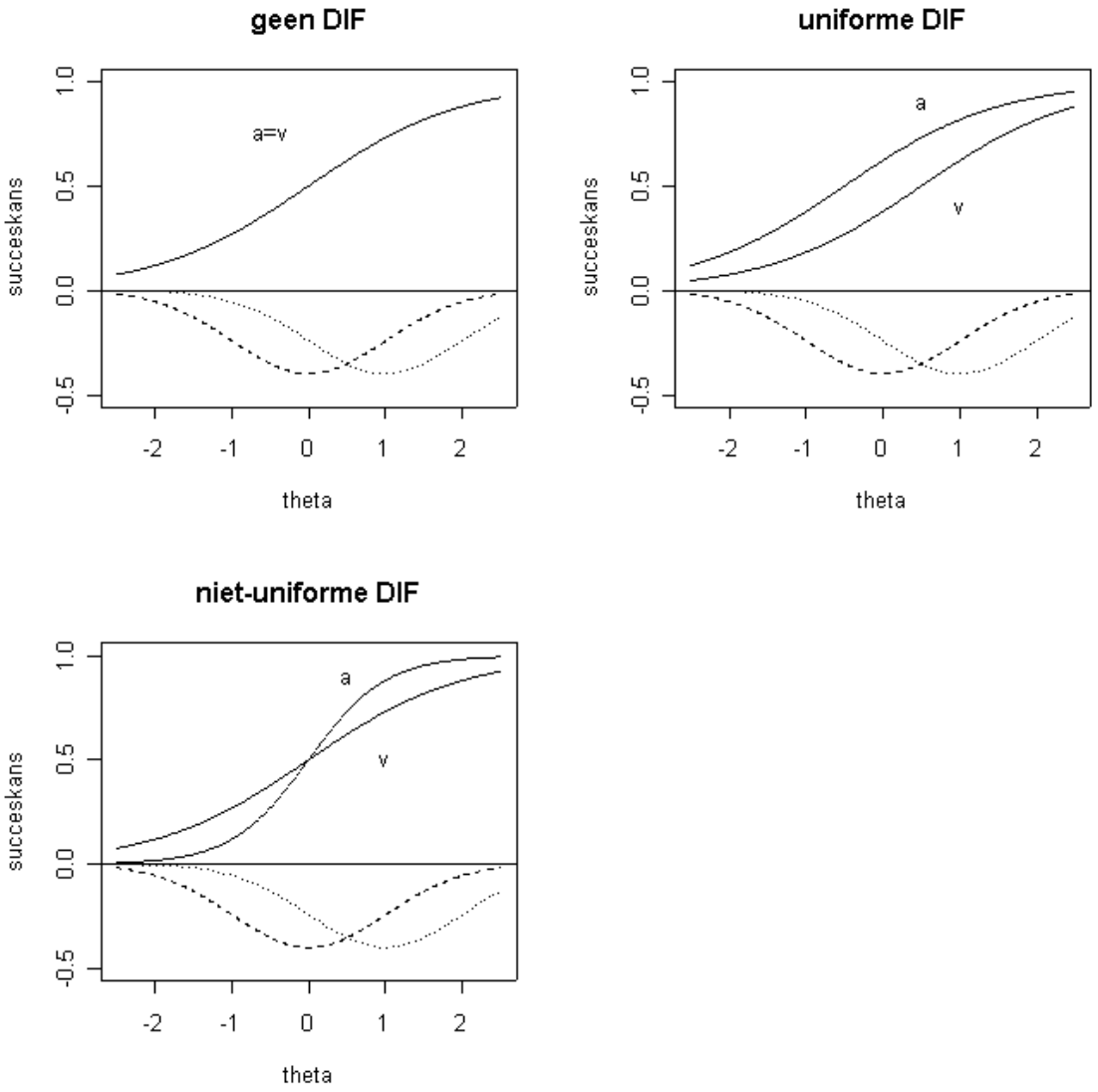
Het linker-boven paneel in Figuur 3.4 toont de itemresponsfunctie van een item dat geen DIF vertoont. We merken op dat beide groepen wel een andere vaardigheidsverdeling kunnen hebben. In Figuur 3.4 hebben Vlamingen gemiddeld een hoger vaardigheidsniveau dan allochtonen en is de spreiding van de vaardigheid dezelfde in de twee groepen.

Om te onderzoeken of een item DIF vertoont moet men onderzoeken of de itemresponsfuncties in beide groepen een verschillend verloop hebben. Dit kan gebeuren door statistisch te testen of de itemparameters verschillen tussen groepen. Om deze statistische tests uit te voeren is het echter cruciaal dat de itemparameters van de twee groepen op dezelfde schaal geplaatst worden zodat het al dan niet optreden van DIF onderscheiden wordt van het feit dat de verschillende groepen mogelijks een verschillende vaardigheidsverdeling hebben. Het calibreren van de itemparameters in beide groepen op een gemeenschappelijke schaal is slechts mogelijk als men een gemeenschappelijke vergelijkingsbasis veronderstelt voor de twee groepen. In psychometrisch onderzoek naar DIF worden hiervoor verschillende methoden gebruikt.

Een eerste methode is te veronderstellen dat het gemiddelde van de moeilijkheidsgraden en het geometrisch gemiddelde van de discriminatiegraden gelijk is in de twee populaties waar de personen van de twee groepen uit afkomstig zijn. We zullen deze methode verder aanduiden als de *methode van gelijke populatiegemiddelden*. Deze werkwijze komt er in de praktijk op neer dat men in een eerste stap de parameters van het itemresponsmodel bepaalt op basis van de testgegevens van elke groep en dat men in een tweede stap de parameters van de tweede groep transformeert zodat de gemiddelde moeilijkheidsgraden en het geometrisch gemiddelde van de discriminatiegraden hetzelfde is in de twee groepen. In een derde stap kan men per item nagaan of DIF optreedt door te testen of de (getransformeerde) moeilijkheidsgraden en discriminatiegraden verschillen in de twee groepen.

Een tweede methode is te veronderstellen dat de itemresponsfuncties voor een bepaalde verzameling van ankeritems een gelijk verloop hebben in de twee groepen. Deze methode wordt verder aangeduid als de *ankermethode*. Het anker laat toe om de testgegevens van de twee groepen samen te analyseren en verschaft een gemeenschappelijke vergelijkingsbasis om de itemparameters van beide groepen op een gemeenschappelijke schaal te plaatsen. Voor niet-anker items kan men in een volgende fase bepalen of DIF optreedt door te testen of moeilijkheidsgraden of discriminatiegraden verschillen in de twee groepen.

Aangezien men op voorhand meestal niet weet welke items zuiver zijn kiest men het anker soms empirisch op basis van voorafgaande analyses (bijvoorbeeld alle items die op zich DIF vertonen als alle andere items als anker gebruikt worden). Een eenvoudige strategie is ankeritems te kiezen die volgens de methode van gelijke populatiegemiddelden sterk op elkaar gelijkende itemresponsfuncties hebben (en dus geen DIF vertonen).



Figuur 3.4 Itemresponsfunctie voor allochtonen (a) en Vlamingen (v) voor een item zonder DIF, met uniforme DIF en met niet-uniforme DIF. De vaardigheidsverdelingen voor allochtonen (- - -) en Vlamingen (….) zijn omgekeerd weergegeven in de onderste helft van elke figuur.

Om het modelleren van DIF formeel te beschrijven veronderstellen we dat we beschikken over een test met M ankeritems (i=1,..,M) en I-M items die moeten onderzocht worden op DIF (i=M+1,..,I). Nemen we verder aan dat het 3PL met gelijke raadparameters in de twee groepen de itemresponsgegevens goed beschrijft (γ_i voor alle items hetzelfde in elke groep) dan ziet het model voor de ankeritems er als volgt uit:

$$\Pr(Y_{pi}=1|\theta_p,\alpha_i,\beta_i,\gamma_i,z_p)=\gamma_i+(1-\gamma_i)\frac{\exp[\alpha_i(\theta_p-\beta_i)]}{1+\exp[\alpha_i(\theta_p-\beta_i)]} \qquad (3.2)$$

en het model voor items die onderzocht dienen te worden voor DIF ziet er als volgt uit:

$$\Pr(Y_{pi}=1|\theta_p,\alpha_i,\beta_i,\gamma_i,\varepsilon_i,\xi_i,z_p)=\gamma_i+(1-\gamma_i)\frac{\exp\left[(\alpha_i+z_p\varepsilon_i)(\theta_p-(\beta_i+z_p\xi_i))\right]}{1+\exp\left[(\alpha_i+z_p\varepsilon_i)(\theta_p-(\beta_i+z_p\xi_i))\right]}. \qquad (3.3)$$

Zoals blijkt uit formules (3.2) en (3.3) zijn voor de ankeritems de succeskansen dezelfde in beide groepen, terwijl voor de niet-anker items specifieke succeskansen gelden voor elke groep. Voor Vlamingen (Z=0) gelden moeilijkheidsgraden β_i en discriminatiegraden α_i en voor allochtonen (Z=1) gelden moeilijkheidsgraden $\beta_i+\xi_i$ en discriminatiegraden $\alpha_i+\varepsilon_i$. De parameters ξ_i vormen het verschil tussen moeilijkheidsgraden in beide groepen en de parameters ε_i vormen het verschil tussen discriminatiewaarden in beide groepen. Om te onderzoeken of DIF optreedt in item i moet men statistisch testen of de DIF-gerelateerde parameters ξ_i en ε_i verschillen van 0.

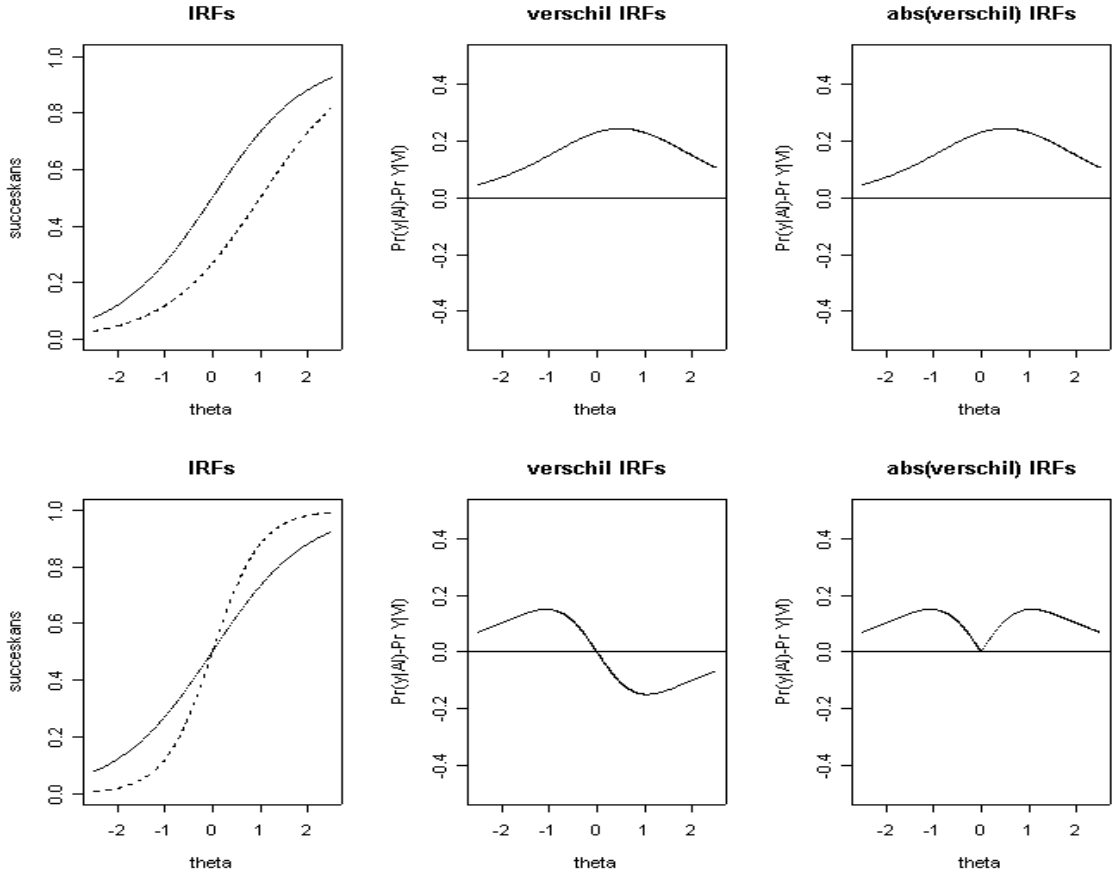
In formules (3.2)-(3.3) veronderstellen we ook dat de latente variabele θ een verschillende verdeling kan hebben naargelang van de groep, namelijk $\theta \sim N(0,1)$ voor Vlamingen en $\theta \sim N(\mu,\sigma^2)$ voor allochtonen. We merken hierbij op dat de parameters van de normale verdeling bij Vlamingen vastgelegd worden op arbitraire waarden (in dit geval gemiddelde gelijk aan 0 en variantie gelijk aan 1) om het nulpunt en de eenheid van de latente schaal te bepalen. De parameter μ geeft de positie van de gemiddelde allochtoon op de schaal terwijl de positie van de gemiddelde autochtoon 0 is. De verdeling van de vaardigheid θ kan dus anders zijn naargelang van de groep maar dit staat los van het feit of er al dan niet DIF optreedt in sommige items. DIF gaat immers over verschillen in succeskansen voor allochtonen en Vlamingen met dezelfde positie op de schaal.

Er kunnen verschillende types DIF onderscheiden worden (zie Mellenbergh, 1982): Men spreekt van uniforme DIF in een item als alleen de moeilijkheidsgraad verschilt in beide groepen ($\varepsilon_i=0$ en $\xi_i\neq0$). Het rechter-boven paneel van Figuur 3.4 toont een item dat uniforme DIF vertoont. Unidirectionele DIF (Shealy & Stout, 1993a, 1993b) treedt op als uniforme DIF voor alle items in het voordeel van dezelfde groep is (maar niet noodzakelijk even sterk). Bij niet-uniforme DIF verschilt de discriminatiewaarde in beide groepen ($\varepsilon_i\neq0$) en mogelijks ook de moeilijkheidsgraad (cf. het linker-onder paneel in Figuur 3.4). Geen DIF impliceert tenslotte dat zowel de moeilijkheidsgraad als de discriminatiegraad van het item niet significant verschillen in beide groepen ($\varepsilon_i=\delta_i=0$) (cf. linker-boven paneel in Figuur 3.4).

Omdat een statistisch significant verschil in parameterwaarden bij grote steekproeven niet noodzakelijk praktisch significant is, is het van belang om het effect van DIF op de succeskans van verschillende groepen in kaart te brengen. Men kan dit bijvoorbeeld doen door het verschil tussen de itemresponsfuncties te visualiseren of door de verdeling van de absolute waarde van de verschillen tussen de itemresponsfuncties over het bereik van de latente schaal te visualiseren of samenvattend te beschrijven. Figuur 3.5 visualiseert voor items met uniforme en niet-uniforme DIF het verschil tussen itemresponsfuncties en de absolute waarde van het verschil tussen itemresponsfuncties.

Het bovenste paneel van Figuur 3.5 toont de grootte van het verschil tussen itemresponsfuncties en van de absolute waarde van het verschil tussen itemresponsfuncties met uniforme DIF. Merk op dat het verschil en het absolute verschil een identiek verloop kennen in geval van uniforme DIF. Om de (absolute) waarde van het verschil samenvattend te beschrijven kunnen we de percentielen van de gediscretiseerde verdeling rapporteren. Meerbepaald blijkt dat het absolute verschil varieert tussen .05 en .24 (op de schaal van de succeskans) en dat de mediaan van de verschilscores gelijk is aan .17. Het 95% betrouwbaarheidsinterval van de verschilscores is (.05,.24).

Het onderste paneel van Figuur 3.5 toont het verschil en het absolute verschil tussen itemresponsfuncties met niet-uniforme DIF. In tegenstelling tot het geval van uniforme DIF zijn deze functies niet langer identiek. Om de grootte en de ernst van de DIF samen te vatten is het nu beter om de verdeling van absolute verschilscores te rapporteren in plaats van de verdeling van de gewone verschilscores. De verdeling van de gemiddelde verschilscores is immers gelijk aan 0 omdat positieve DIF in het begin van de schaal en negatieve DIF aan het eind van de schaal elkaar opheffen. De verdeling van de absolute verschilscores daarentegen varieert tussen 0 en .15 en heeft een mediaan gelijk aan .12. Het 95% betrouwbaarheidsinterval van de absolute verschilscores is (.02,.15).



Figuur 3.5 Itemresponsfuncties, verschil tussen itemresponsfuncties en absolute waarde van het verschil tussen itemresponsfuncties voor items met uniforme DIF (bovenste paneel) en niet-uniforme DIF (onderste paneel).

De concrete strategie die gebruikt wordt voor het modelleren van DIF in dit rapport is dezelfde voor alle datasets. Deze strategie bestaat uit drie stappen:

(1) De parameters van het itemresponsmodel worden geschat voor elke groep apart en ze worden op een gemeenschappelijke schaal geplaatst op basis van de methode van gelijke populatiegemiddelden. Omdat we beschikken over relatief kleine steekproeven van autochtonen en allochtonen worden de raadparameters gelijk gesteld aan 0 en wordt verondersteld dat alle items gelijke discriminatiegraden hebben. Er wordt dus per item alleen een verschillende moeilijkheidsgraad geschat. Dit strengere model is in de literatuur ook bekend als het "Raschmodel".

(2) Er wordt aan de hand van statistische tests voor elk item nagegaan of er DIF optreedt in de moeilijkheidsgraad ($\xi_i \neq 0$). Aangezien verondersteld wordt dat de discriminatiegraden gelijk zijn over groepen, zullen we alleen uniforme DIF opsporen.

Om de parameters van het Raschmodel te schatten maken we gebruik van algoritmen voor Bayesiaanse data analyse die geïmplementeerd zijn in de software WINBUGS (Spiegelhalter, Thomas, & Best, 1999). Deze aanpak heeft als voordeel dat de standaardfouten van de DIF parameters ook nauwkeurig kunnen geschat worden bij relatief kleine steekproeven waardoor de bijhorende DIF tests ook nauwkeurig zijn.

3.3 Verklaren van DIF

Nadat voor een bepaalde test onderzocht is voor welke items DIF optreedt, rijst de vraag hoe men de vastgestelde DIF kan verklaren. Dit kan door te onderzoeken of DIF gemodelleerd kan worden als een functie van itemkenmerken. De itemkenmerken kunnen bijvoorbeeld het resultaat zijn van een cognitieve analyse van de items. Dit wordt in het onderstaand kader geïllustreerd voor een item dat transitief redeneren meet. In deze test moet men op basis van gegeven relaties tussen A en B en tussen A en C de relatie tussen A en C afleiden. Mogelijke itemkenmerken die de moeilijkheidsgraad van een item bepalen zijn het aantal proposities, het aantal en het type van cognitieve operaties die men moet uitvoeren om tot een oplossing te komen.

Het verklaren van DIF wil zeggen dat men verschillen in moeilijkheidsgraden tussen de groepen probeert te verklaren op basis van itemkenmerken. De itemkenmerken worden aangeduid als variabelen F_1 t.e.m. F_K. De score van item i op kenmerk k is gelijk aan F_{ik}. Men kan eventueel nog een stap verder gaan dan het louter modelleren van verschillen in moeilijkheidsgraden en moeilijkheidsgraden zelf in beide groepen trachten te modelleren als een functie van itemkenmerken. Als itemkenmerken naargelang van de groep een andere bijdrage hebben aan de moeilijkheidgraad van items, dan spreken we van differential feature functioning (DFF). De verschillende modellen voor DFF kunnen beschreven worden door in de exponent van formule (3.3) bepaalde parameters te modelleren als een functie van itemkenmerken:

Voorbeeld: Logisch redeneren

voorbeelditem: B rijdt niet trager dan A, B rijdt niet sneller dan C
- De relatie tussen A en C is niet te bepalen
- A is trager dan C
- A is sneller dan C
- A is niet sneller dan C
- A is niet trager dan C

5 itemkenmerken:

- F_1 = Het aantal premissen dat gegeven is bij een bepaald item. Bijvoorbeeld, in bovenstaand item zijn er twee premissen “B rijdt niet trager dan A” en “B rijdt niet sneller dan C”.
- F_2 = Het aantal keer dat de relatie "kleiner dan of gelijk aan" voorkomt in de gegeven premissen. Premissen kunnen een strikte orde-relatie uitdrukken tussen objecten (“A is kleiner dan B”) of een niet strikte-orde relatie (“A is kleiner dan of gelijk aan B” of anders geformuleerd “A is niet groter dan B”). In het voorbeeld-item drukken beide premissen een niet-strikte orde relatie uit.
- F_3 = Het aantal omwisselingen van premissen dat nodig is om tot een correct antwoord te komen. Er wordt verondersteld dat proefpersonen van de gekende transitieve regel “(A is kleiner dan B) en (B is kleiner dan C) dus (A is kleiner dan C) ” gebruik maken. Om deze regel te kunnen toepassen moeten de premissen in de juiste volgorde staan.
- F_4 = Het aantal omwisseling binnen premissen dat nodig is om tot een correct antwoord te komen. Bij het toepassen van de transitieve regel is het soms nodig om binnen een premisse de volgorde van de objecten te veranderen. Bijvoorbeeld, “B is kleiner dan A” kan geherformuleerd worden als “A is groter dan B”.
- F_5 = Het juiste antwoordalternatief bevat de relatie “kleiner dan of gelijk aan” (F_5=1) of niet (F_5=0).

Tabel 3.1 beschrijft de modellen die het meest van belang zijn in het kader van dit rapport, namelijk modellen om moeilijkheidsgraden of een verschil in moeilijkheidsgraden te verklaren op basis van itemkenmerken. Model 1 gebruikt aparte parameters voor de moeilijkheidsgraden in elke groep en heeft dus geen verklarende waarde. Model 2 modelleert het verschil tussen de moeilijkheidsgraden in de twee groepen als een functie van itemkenmerken en tracht dus een verklaring te geven voor DIF in de moeilijkheidsgraden. Model 3 gaat nog een stap verder en modelleert de moeilijkheidsgraden in elke groep als een functie van itemkenmerken. Dit model probeert op basis van een inhoudelijke theorie over de cognitieve operaties die nodig zijn om het item op te lossen een inzicht te geven in de algemene moeilijkheid van de items en hoe dit verschilt in de twee groepen.

Tabel 3.1 Modellen voor DFF

Model	Exponent van formule (2.3)
1.Geen verklaring	$[\theta_p-(\beta_i+z_p\xi_i)]$
2.Verklaren ξ_I	$[\theta_p-(\beta_i+z_p\sum_k \eta_k F_{ik})]$
3.Verklaren β_i en ξ_i	$[\theta_p-(\tau_0+\sum_k\tau_k F_{ik}+z_p\sum_k \eta_k F_{ik})]$

Om praktisch te evalueren in welke mate DIF parameters of moeilijkheidsgraden in elke groep kunnen verklaard worden op basis van objectieve itemkenmerken zullen we de parameters (η,τ) van de modellen in Tabel 3.1 niet rechtstreeks schatten maar zullen we ze benaderen via eenvoudige multiple regressie op de geschatte DIF parameters en moeilijkheidsgraden. Bij deze analyses fungeren verschillen in moeilijkheidsgraden (DIF parameters) of moeilijkheidsgraden van een bepaalde groep als criterium en itemkenmerken als predictoren.

3.4 Effect van DIF in individuele items op de testscore

De resultaten van de DIF analyse vertellen ons welke individuele items DIF vertonen. Daarnaast is het interessant om na te gaan in welke mate de DIF in alle individuele items samen de prestatie van personen uit verschillende groepen met dezelfde positie op de schaal differentieel beïnvloedt. Om de prestatie van een persoon op een test te evalueren kan men verschillende maten gebruiken. Een eerste veel gebruikte maat is het aantal juiste antwoorden of de somscore. Deze maat wordt gebruikt bij de datasets van VDAB. Bij de MCT-M wordt er namelijk niet gecorrigeerd voor raden. Een tweede populaire maat die gangbaar is bij tests met meerkeuze-items en die o.a. gebruikt wordt door SELOR en door ABL bij de tests die in dit rapport bestudeerd worden is de somscore die gecorrigeerd is voor het aantal foute antwoorden dat een persoon heeft. We zullen in wat volgt bespreken hoe men het effect van DIF in individuele items kan nagaan op de somscore van personen uit verschillende groepen met dezelfde θ.

3.4.1 Somscore

Omdat het itemresponsmodel veronderstelt dat de antwoorden van een persoon op verschillende items onafhankelijk tot stand komen, gegeven zijn/haar positie op de schaal, is het eenvoudig om voor elk punt op de schaal uit te rekenen wat de somscore op de test is en om een 95% betrouwbaarheidsinterval af te bakenen rond deze verwachte somscore. Definiëren we de variabele $S_{pz}=\sum_i Y_{pi}$ als de som van de juiste antwoorden van een persoon uit groep z met positie θ_p. Dan is de verwachte waarde van S_{pz}

$E(S_{pz})=\sum_i Pr(Y_{pi}=1|\theta_p,z_p)$

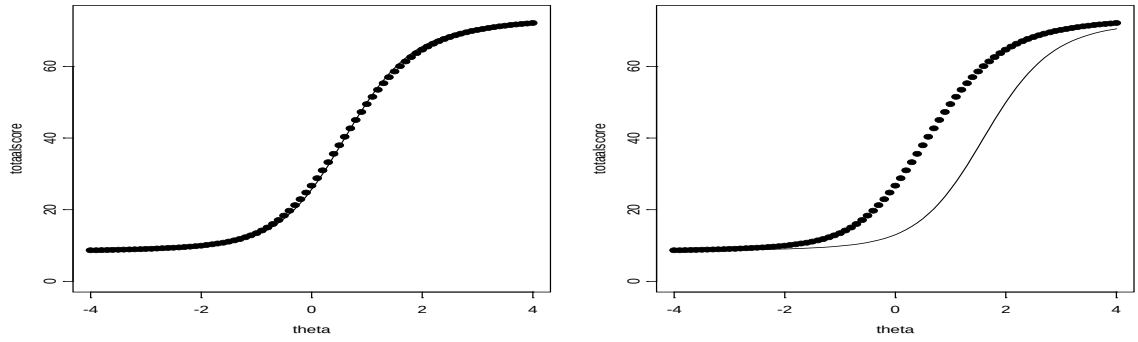
en dan is de variantie van S_{pz}

$VAR(S_{pz}) = \sum_i Pr(Y_{pi}=1|\theta_p,z_p)\,[1-Pr(Y_{pi}=1|\theta_p,z_p)]$.

Zowel de verwachte somscore als de variantie van de somscore voor een bepaald punt op de schaal zijn dus een eenvoudige functie van de succeskansen op de verschillende items van de test.

Een vergelijking van de verwachte-somscore curve voor elke groep toont hoe de DIF in individuele items een mogelijks verschillende invloed heeft op de verwachte somscore in elke groep. Merk op dat de verdeling van de latente variabele in elk van de groepen in principe een verschillend gemiddelde kan hebben en dat de figuur dus enkel het effect van DIF toont. De onderstaande Figuur 3.6 toont de verwachte somscores voor allochtonen (dunne lijn) en Vlamingen (dikke punten) op twee tests. In de linkerfiguur is er bijna geen differentieel effect op de testscore terwijl in de rechterfiguur dit wel het geval is. Merk op dat de linkerfiguur niet noodzakelijk impliceert dat er weinig DIF is in individuele items. Het zou immers kunnen dat een aantal items DIF vertonen in het voordeel van de allochtonen en dat een aantal andere items DIF vertonen in het voordeel van de Vlamingen. Tegengestelde DIF-effecten in individuele items heffen elkaar dan op zodat de DIF in individuele items niet zichtbaar is in de somscore. Het is evenwel belangrijk te beseffen dat eenzelfde somscore in een dergelijk geval iets anders kan betekenen bij allochtonen en Vlamingen omdat eenzelfde somscore kan tot stand komen door succes op verschillende verzamelingen van items die elk een tegengestelde DIF vertonen.

Het is belangrijk op te merken dat de in dit onderzoek gebruikte calibratiemethode (parameters van twee groepen op een gemeenschappelijke schaal plaatsen via assumptie van gelijke populatiegemiddelden) meestal niet zal leiden tot verschillende verwachte-somscore curven in verschillende groepen. De reden hiervoor is dat de assumptie van gelijke populatiegemiddelden het hoofdeffect zo kiest dat de aanwezige DIF zowel positief als negatief is. Bijgevolg zullen deze tegengestelde DIF effecten elkaar meestal compenseren op het niveau van de verwachte somscore-curve.



Figuur 3.6: Een fictief voorbeeld van de verwachte somscores voor allochtonen (dunne lijn) en Vlamingen (dikke punten) op twee tests.

Hoofdstuk 4: DIF-analyses MCT-M

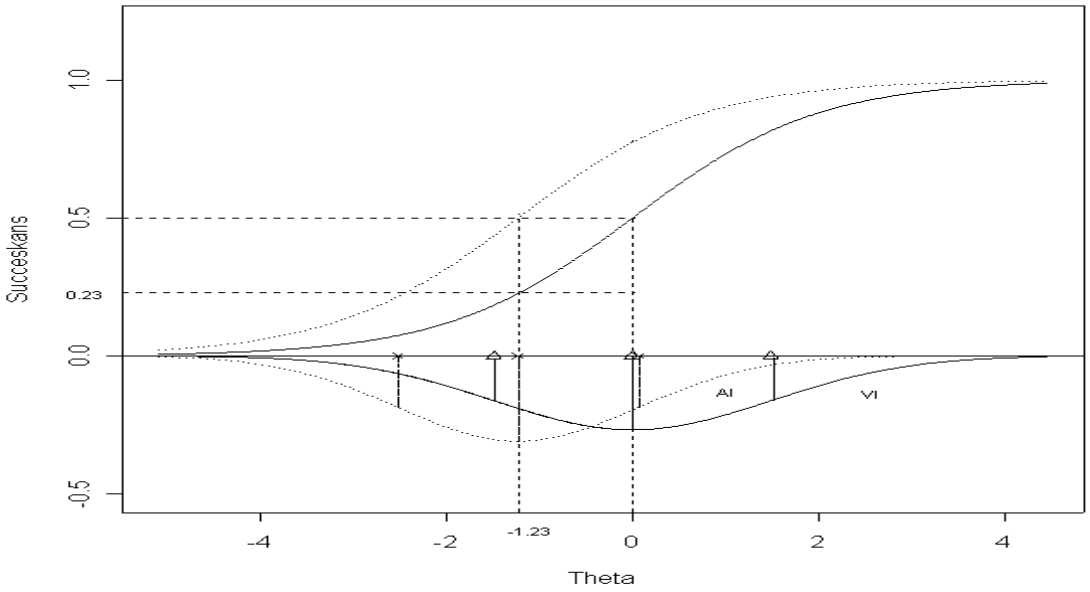
In dit hoofdstuk analyseren we de subtests van de Multiculturele Capaciteitentest (MCT-M). Elke analyse bevat 3 stappen: (1) onderzoeken of individuele items voor Vlamingen versus allochtonen DIF vertonen in de moeilijkheidsgraad (uniforme DIF), (2) als er sprake is van uniforme DIF, deze verklaren in functie van itemkenmerken, (3) onderzoeken wat het effect is van DIF in individuele items op de testscore.

4.1 CIJFERREEKSEN

4.1.1 Modelleren van DIF

Na schatting van het Raschmodel op elke groep worden de itemparameters op dezelfde schaal geplaatst met de methode van gelijke populatiegemiddelden. Het gemiddelde van de verdeling van de latente variabele bij Vlamingen en allochtonen wordt op voorhand vastgelegd op 0. De standaarddeviaties van θ (σ_{vl} en σ_{all}) worden geschat en kunnen dus verschillen in beide groepen. Deze blijken respectievelijk 1.49 en 1.29 te zijn. Wanneer we deze latente variabelen op één schaal plaatsen, geldt voor Vlamingen $\theta \sim N(0,1.49^2)$ en voor allochtonen $\theta \sim N(-1.23,1.29^2)$. We stellen vast dat Vlamingen gemiddeld beter presteren op de test dan allochtonen (μ=-1.23, $p<.01$). Het belang hiervan voor de succeskansen van Vlamingen en allochtonen die gemiddeld presteren kan als volgt geïllustreerd worden: Stel dat voor een bepaald item zowel de moeilijkheidsgraad als de DIF-parameter gelijk is aan 0. Dan heeft de gemiddelde Vlaming ($\theta = 0$) een succeskans van 0.50, terwijl de gemiddelde allochtoon ($\theta = -1.23$) een succeskans van 0.23 heeft. Een gemiddelde Vlaming heeft dus twee keer meer kans om dit item juist op te lossen dan een gemiddelde allochtoon.

In figuur 4.1 vindt u een illustratie van bovenstaand voorbeeld. In deze figuur wordt de verdeling van theta (vaardigheid) voor Vlamingen en allochtonen weergegeven door de curves onder de horizontale lijn. Per groep wordt het gemiddelde van de vaardigheidsverdeling en de percentielpunten 25 en 75 aangeduid met een pijl (Δ voor Vlamingen en X voor allochtonen). Daarnaast wordt per groep het item met gemiddelde moeilijkheidsgraad weergegeven (boven de X-as). Een item met gemiddelde moeilijkheid voor Vlamingen is het item waar de gemiddelde vlaming 50 % kans heeft om het correct op te lossen (__). Een item met gemiddelde moeilijkheid voor allochtonen is een item waar de gemiddelde allochtoon 50 % kans heeft om het correct op te lossen (…). In de figuur is aangeduid dat de gemiddelde allochtoon 23% kans heeft op een item met moeilijkheidsgraad 0. We zien ook dat de 25% beste allochtonen even goed presteren als de gemiddelde Vlaming in de steekproef.
Tabel 4.1 geeft een overzicht van de geschatte itemparameters voor de Vlamingen (β) en van de DIF die optreedt in de moeilijkheidsgraden (ξ). De ξ parameters geven aan hoeveel moeilijker een item is voor allochtonen dan voor Vlamingen. Daarnaast geeft Tabel 4.1 per item een overzicht van de praktische significantie van de DIF aan de hand van de mediaan en het 95% BI van de absolute verschillen tussen de IRFs van Vlamingen en allochtonen.



Figuur 4.1 Verdeling van de latente variabele (Theta) voor Vlamingen (__) en allochtonen (…) en IRFs van unbiased items met gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

Uit Tabel 4.1 blijkt dat er geen significante DIF optreedt. Wanneer er gecorrigeerd wordt voor het hoofdeffect zijn er geen items die significant moeilijker/gemakkelijker zijn voor allochtonen in vergelijking met Vlamingen.
Uit de verdeling van de absolute verschillen tussen IRFs van Vlamingen en allochtonen blijkt dat de praktische significantie van de DIF ook eerder beperkt is. De mediaan van de absolute verschillen in succeskansen varieert van .01 tot .05 en het 97.5 percentiel varieert van .00 tot .21. Absolute verschillen in succeskansen zijn dus in het algemeen erg klein (mediaan), maar ze zijn voor enkele items op een relatief klein deel van de schaal wel van belang.

Merk op dat de moeilijkheidsgraden van de items aan het begin van de test dikwijls erg laag zijn en aan het eind van de test erg hoog. Dit wil zeggen dat items aan het begin van de test meestal juist opgelost worden en dat items aan het eind meestal fout of niet opgelost worden. Voor items met een extreme moeilijkheidsgraad is er weinig informatie voorhanden zodat de standaardfouten van de parameterschatting relatief groot zijn. Bijgevolg zijn ook de standaardfouten van de DIF parameters groot waardoor er minder kans is om significant DIF te vinden. Het onderscheidingsvermogen of de statistische power van de tests voor DIF detectie wordt dus op 2 manieren gedrukt: (1) door het feit dat we slechts over relatief kleine steekproeven beschikken, (2) door het feit dat items aan het begin/eind van de test extreem gemakkelijk/moeilijk zijn.

Tabel 4.1 Parameters van de DIF analyse. Mediaan en 95% Betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	SD(β)	ξ	SD(ξ)	Mediaan	95%BI
1	-4,14	.37	-.17	.45	.00	[.00,.04]
2	-4,64	.46	-.10	.56	.00	[.00,.02]
3	-3,90	.36	.20	.42	.00	[.00,.05]
4	-4,63	.46	.89	.51	.01	[.00,.21]
5	-4,43	.43	-.03	.51	.00	[.00,.01]
6	-4,01	.37	-.34	.45	.01	[.00,.08]
7	-4,65	.48	-.09	.56	.00	[.00,.02]
8	-3,89	.35	-.42	.43	.01	[.00,.10]
9	-3,90	.36	-.01	.43	.00	[.00,.00]
10	-3,77	.35	.23	.39	.00	[.00,.06]
11	-3,10	.27	.21	.34	.01	[.00,.05]
12	-1,72	.20	-.45	.26	.05	[.00,.11]
13	-2,21	.22	-.23	.27	.02	[.00,.06]
14	-1,76	.21	.19	.26	.02	[.00,.05]
15	-2,02	.21	-.42	.26	.04	[.00,.11]
16	-1,39	.20	-.28	.25	.03	[.00,.07]
17	-1,07	.18	-.08	.23	.01	[.00,.02]
18	-0,59	.18	.30	.24	.03	[.00,.08]
19	-0,59	.18	.21	.23	.02	[.00,.05]
20	-0,81	.18	.41	.25	.04	[.00,.10]
21	-0,09	.18	.16	.25	.02	[.00,.04]
22	0,52	.18	-.13	.25	.01	[.00,.03]
23	0,84	.19	.10	.26	.01	[.00,.02]
24	0,55	.19	.07	.27	.01	[.00,.02]
25	1,21	.20	.12	.28	.01	[.00,.03]
26	1,23	.20	-.08	.28	.01	[.00,.02]
27	1,05	.20	-.12	.27	.01	[.00,.03]
28	1,74	.21	-.25	.31	.03	[.00,.06]
29	2,11	.22	.43	.38	.03	[.00,.11]
30	2,58	.25	-.31	.36	.02	[.00,.08]

* p<.05; **p<.01

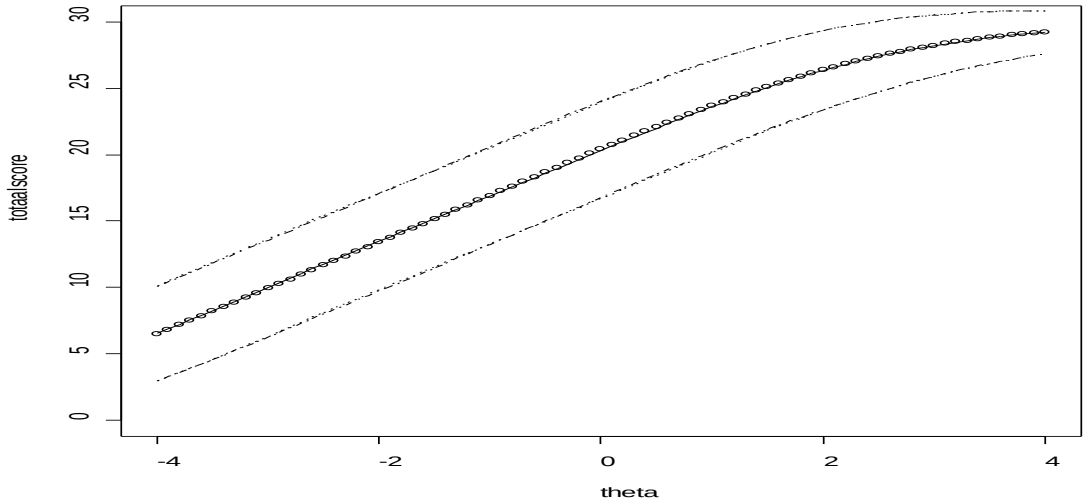
4.1.2 Verklaren van DIF

Er wordt niet naar een verklaring gezocht voor de DIF omdat er geen significante DIF gevonden werd.

4.1.3 Effect van DIF op de testscore

Omdat men in selectieprocedures dikwijls alleen maar gebruik maakt van de somscore die een persoon behaalt op een test, is het van belang om het effect van DIF op de testscore te onderzoeken. Zoals uitgelegd in Hoofdstuk 3 kan dit door per groep de verwachte somscore (en bijbehorend betrouwbaarheidsinterval) te berekenen in functie van θ. Figuur 4.2 beschrijft (uitgaande van de parameterwaarden in Tabel 4.1) de verwachte testscore voor Vlamingen en allochtonen in functie van θ.

Uit de Figuur blijkt dat de verwachte somscore-curves voor beide groepen nauwelijks verschillen. Berekening van bijhorende betrouwbaarheidsintervallen toont dat er inderdaad voor geen enkele waarde van θ een significant verschil is tussen verwachte somscores van Vlamingen en allochtonen.



Figuur 4.2. Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

4.1.4 Conclusie

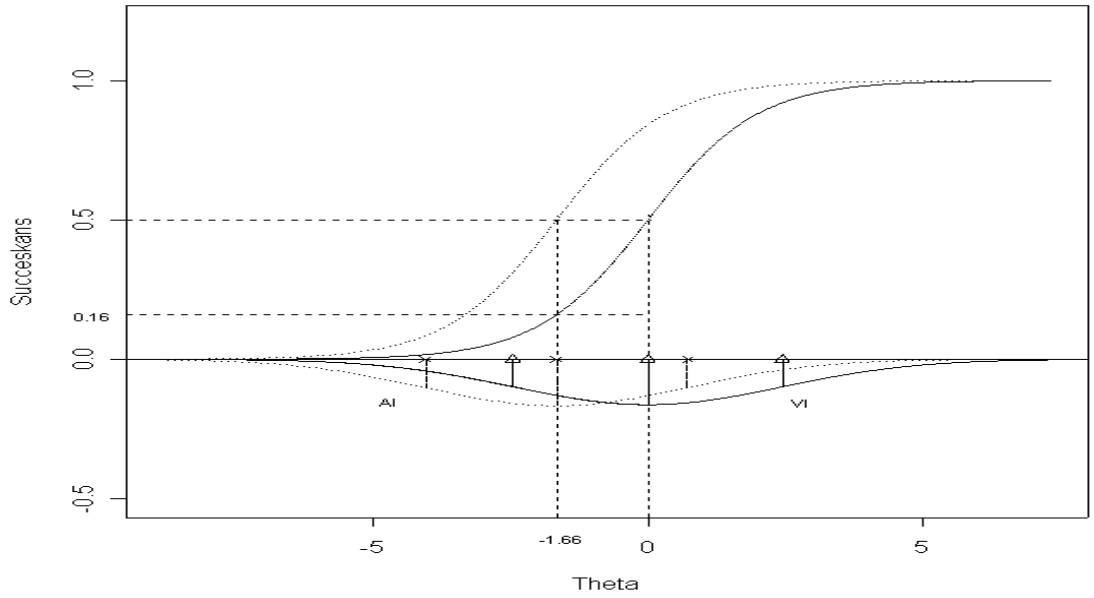
We stellen vast dat, in de huidige steekproef, Vlamingen gemiddeld beter presteren op de subtest cijferreeksen dan allochtonen. Op een vaardigheidsschaal θ waarbij voor Vlamingen geldt dat $\theta \sim N(0,1.49^2)$ is de gemiddelde vaardigheid van allochtonen (uitgedrukt in SDs van de verdeling voor Vlamingen) iets minder dan 1 SD lager (namelijk -1.23). Dit betekent bijvoorbeeld dat een gemiddelde Vlaming twee keer meer kans heeft om een unbiased item met $\beta = 0$ juist op te lossen dan een gemiddelde allochtoon.

Verder stellen we vast dat er bij geen enkel item statistisch significante DIF optreedt. De praktische significantie van de DIF is ook zeer beperkt. Ten slotte blijkt dat de gezamenlijke invloed van DIF in individuele items geen effect heeft op (verwachte) somscores.

4.2 SPIEGELBEELDEN

4.2.1 Modelleren van DIF

De gegevens van Vlamingen en allochtonen werden elk geanalyseerd met het Raschmodel. Daarna werden de parameters op één schaal geplaatst volgens de methode van gelijke populatiegemiddelden. Voor Vlamingen geldt dat $\theta \sim N(0,2.45^2)$ en voor allochtonen
$\theta \sim N(-1.66,2.39^2)$. We stellen dus vast dat in de verzamelde steekproeven Vlamingen gemiddeld beter presteren dan allochtonen (μ=-1.66, p<.01). Ter illustratie toont Figuur 4.3 de vaardigheidsverdeling voor allochtonen en Vlamingen en de itemresponsfuncties van unbiased items met gemiddelde moeilijkheid voor Vlamingen en allochtonen. We kunnen bijvoorbeeld uit de figuur aflezen dat de gemiddelde Vlaming een succeskans van .50 heeft voor een item met β=0 en dat de succeskans voor een gemiddelde allochtoon drie keer lager is, namelijk .16.



Figuur 4.3: Verdeling van de latente variabele voor Vlamingen (__) en allochtonen (…) en IRF van een unbiased item met gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

Tabel 4.2 toont de geschatte itemparameters voor Vlamingen (β) en de geschatte DIF-parameters (ξ). Een positieve ξ wil zeggen dat het item moeilijker is voor allochtonen terwijl een negatieve ξ wil zeggen dat het makkelijker is voor allochtonen. De laatste twee kolommen van de tabel geven informatie over de praktische significantie van de DIF, namelijk, de mediaan en het 95% betrouwbaarheidsinterval van de absolute verschillen tussen IRFs voor Vlamingen en allochtonen. We stellen vast dat ook in deze test geen enkel item significante DIF vertoont op het 5% niveau. De mediaan van de

absolute verschillen tussen IRFs van Vlamingen en allochtonen varieert van .00 tot .06 en heeft een gemiddelde waarde van .02. Het 97.5 percentiel van de absolute verschillen varieert tussen .00 en .14. We kunnen besluiten dat de praktische significantie van de DIF over het algemeen beperkt is.

Tabel 4.2 Parameters van de DIF analyse. Mediaan en 95% Betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	SD(β)	ξ	SD(ξ)	Mediaan	95% BI
1	-2.43	.28	.05	.34	.00	[.00,.01]
2	-1.45	.27	-.07	.34	.01	[.00,.02]
3	-1.41	.26	-.51	.34	.05	[.00,.13]
4	-1.02	.27	-.07	.34	.01	[.00,.02]
5	-1.76	.27	.30	.34	.03	[.00,.07]
6	-1.96	.27	.25	.35	.03	[.00,.06]
7	-1.26	.26	-.14	.34	.01	[.00,.03]
8	-1.50	.27	.10	.34	.01	[.00,.02]
9	-1.26	.26	.02	.33	.00	[.00,.01]
10	-.44	.26	-.35	.34	.04	[.00,.09]
11	-1.77	.28	.03	.35	.00	[.00,.01]
12	-1.15	.27	.26	.36	.03	[.00,.06]
13	-.07	.26	-.37	.34	.04	[.01,.09]
14	-1.49	.27	.37	.34	.04	[.00,.09]
15	-1.19	.26	-.05	.34	.00	[.00,.01]
16	-.87	.26	.26	.35	.03	[.00,.06]
17	-1.29	.27	.30	.35	.03	[.00,.07]
18	-.90	.27	.07	.34	.01	[.00,.02]
19	-.23	.25	-.01	.34	.00	[.00,.00]
20	-.31	.26	-.12	.35	.01	[.00,.03]
21	.06	.27	-.10	.36	.01	[.00,.03]
22	-.47	.26	.13	.34	.01	[.00,.03]
23	.28	.26	.20	.35	.02	[.00,.05]
24	.31	.25	.06	.33	.01	[.00,.02]
25	.74	.26	.31	.36	.03	[.00,.08]
26	.37	.26	-.17	.34	.02	[.00,.04]
27	1.69	.27	-.56	.38	.06	[.00,.14]
28	1.12	.26	.31	.37	.03	[.00,.08]
29	1.32	.26	-.27	.36	.03	[.00,.07]
30	1.25	.27	-.25	.36	.03	[.00,.06]

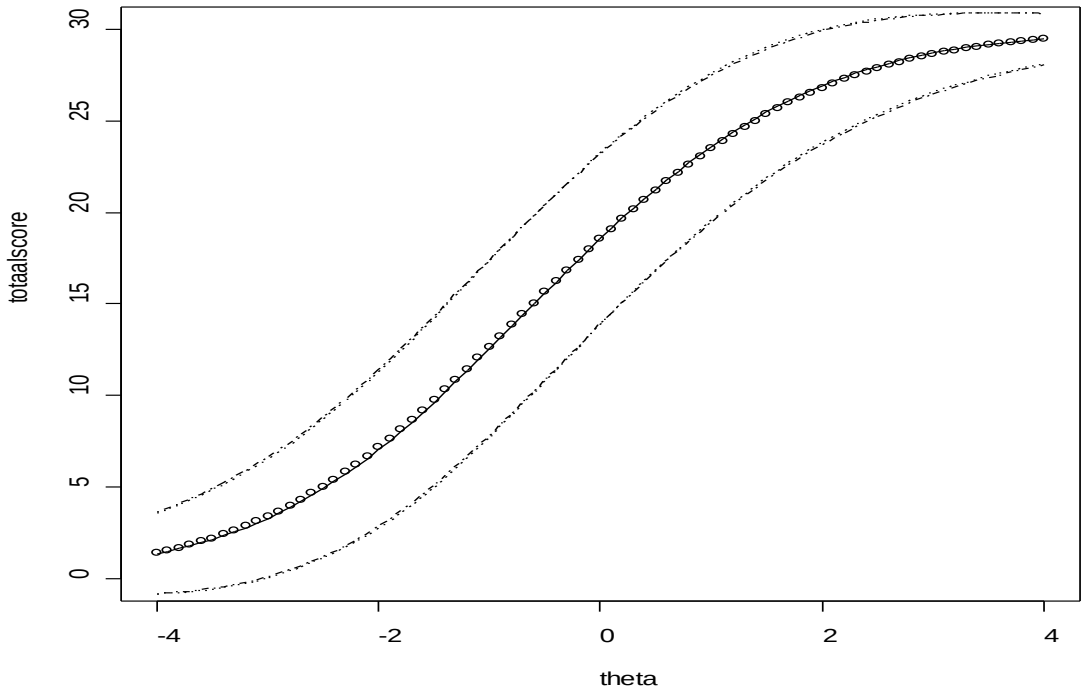
* p<.05; **p<.01

4.2.2 Verklaren van DIF

Aangezien er geen betekenisvolle DIF gevonden werd, is er geen verklaring nodig.

4.2.3 Effect van DIF op de testscores

Om het belang van DIF op de testscores na te gaan, wordt in Figuur 4.4 de verwachte somscores per groep (en bijhorend betrouwbaarheidsinterval) weergegeven. De verwachte somscore-curves verschillen niet significant voor beide groepen.



Figuur 4.4 Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

4.2.4 Conclusie

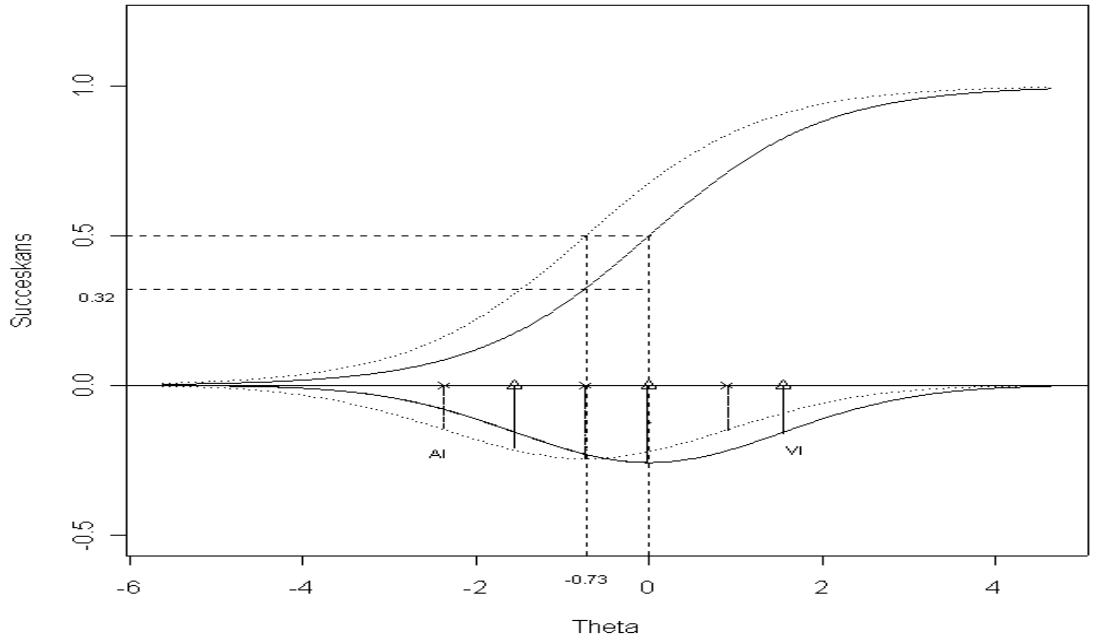
We stellen vast dat voor de huidige gegevens de gemiddelde Vlaming 3 keer beter presteert dan de gemiddelde allochtoon. Verder blijkt dat geen van de 30 items statistisch significante DIF vertoont. De praktische significantie van de DIF is ook zeer beperkt (de mediaan van absolute verschillen tussen IRFs van Vlamingen en allochtonen heeft een gemiddelde waarde van .02). Er is geen verschil in verwachte somscores voor Vlamingen versus allochtonen.

4.3 KOMPONENTEN

4.3.1 Modelleren van DIF

Het Raschmodel wordt per groep geschat en vervolgens worden de parameters op één schaal geplaatst met de methode van gelijke populatie gemiddelden.

Voor Vlamingen geldt dat $\theta \sim N(0,1.55^2)$ en voor allochtonen $\theta \sim N(-.73,1.63^2)$. We stellen dus vast dat in de verzamelde steekproeven Vlamingen gemiddeld beter presteren dan allochtonen
($\mu=-.73$, $p<.01$). Ter illustratie toont Figuur 4.5 de vaardigheidsverdeling voor allochtonen en Vlamingen en de itemresponsfuncties van unbiased items met gemiddelde moeilijkheid voor Vlamingen en allochtonen. We kunnen bijvoorbeeld uit de figuur aflezen dat de gemiddelde Vlaming een succeskans van .50 heeft voor een item met $\beta=0$ en dat de succeskans voor een gemiddelde allochtoon .32 is.



Figuur 4.5 Verdeling van de latente variabele voor Vlamingen (__) en allochtonen (…) en IRFs van unbiased items met gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

Tabel 4.3 geeft een overzicht van de geschatte itemparameters bij Vlamingen (β) en van de DIF in de moeilijkheidsgraden (ξ). De twee laatste kolommen van de Tabel bevatten informatie over de praktische significantie van de DIF, meerbepaald, de mediaan en het 95% betrouwbaarheidsinterval van de verdeling van de absolute verschillen tussen de IRFs van Vlamingen en allochtonen.

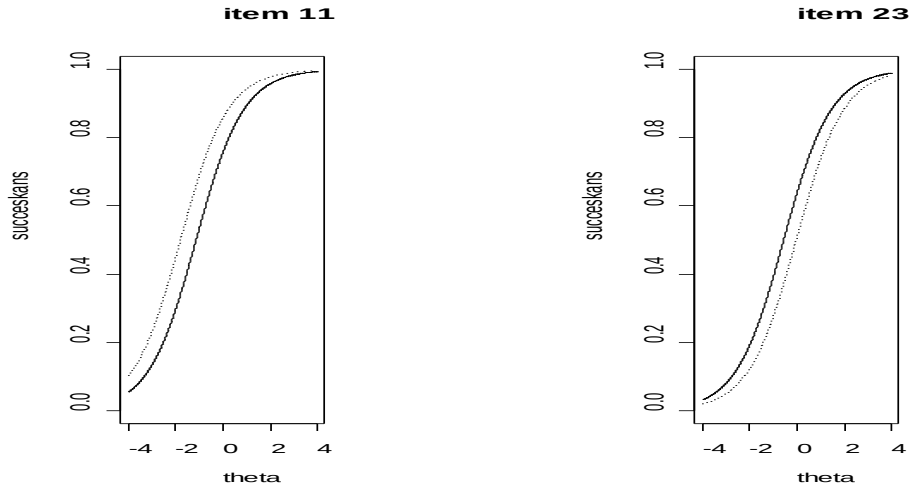
Tabel 4.3 Parameters van de DIF analyse. Mediaan en 95% Betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	SD(β)	ξ	SD(ξ)	Mediaan	95% BI
1	-3.26	.28	-.36	.37	.01	[.00,.09]
2	-4.52	.44	.31	.52	.00	[.00,.07]
3	-2.95	.27	-.27	.33	.01	[.00,.07]
4	-4.21	.37	.23	.45	.00	[.00,.06]
5	-3.26	.29	-.66	.39	.02	[.00,.16]
6	-3.66	.34	.16	.41	.00	[.00,.04]
7	-3.41	.28	-.01	.36	.00	[.00,.00]
8	-3.50	.30	.14	.38	.00	[.00,.03]
9	-2.79	.25	-.33	.32	.02	[.00,.08]
10	-3.4	.29	-.36	.37	.01	[.00,.09]
11	-1.14	.20	-.66**	.26	.07	[.00,.16]
12	-2.95	.26	-.38	.35	.02	[.00,.10]
13	-2.35	.23	-.09	.29	.01	[.00,.02]
14	-.63	.18	-.08	.25	.01	[.00,.02]
15	-1.91	.21	-.06	.28	.01	[.00,.02]
16	-2.18	.23	-.11	.30	.01	[.00,.03]
17	-1.86	.22	.30	.28	.03	[.00,.07]
18	-2.28	.22	.29	.29	.03	[.00,.07]
19	-1.91	.21	.31	.28	.03	[.00,.08]
20	-.40	.18	.15	.26	.02	[.00,.04]
21	-1.09	.19	-.04	.26	.00	[.00,.01]
22	-1.17	.20	.35	.26	.04	[.00,.09]
23	-.57	.18	.53*	.24	.06	[.01,.13]
24	-.42	.19	.35	.25	.04	[.01,.09]
25	.05	.18	.22	.26	.02	[.00,.05]
26	.34	.19	.04	.26	.00	[.00,.01]
27	.37	.19	.02	.26	.00	[.00,.00]
28	1.24	.20	.30	.29	.03	[.00,.07]
29	1.65	.20	.07	.30	.01	[.00,.02]
30	2.02	.21	-.32	.29	.03	[.00,.08]

* p<.05; **p<.01

Uit Tabel 4.3 blijkt dat er slechts bij zeer weinig items significante DIF optreedt (enkel bij item 11 in het voordeel van de allochtonen en bij item 23 in het voordeel van de Vlamingen). Ter illustratie worden items met statistisch significante DIF ook weergegeven in Figuur 4.6. De praktische significantie van de DIF is zeer beperkt. De

mediaan van de absolute verschillen tussen IRFs is gemiddeld .02 en is nooit groter dan .07. Het 97.5 percentiel van de absolute verschillen in succeskansen is in het algemeen tamelijk klein. Het varieert bij alle items van .00 tot .16. We merken nog op dat de items van de test over het algemeen redelijk gemakkelijk zijn voor de onderzochte populatie van personen. Hierdoor zullen de standaardfouten van de geschatte parameters relatief groot zijn en is het moeilijker om significante DIF te vinden.



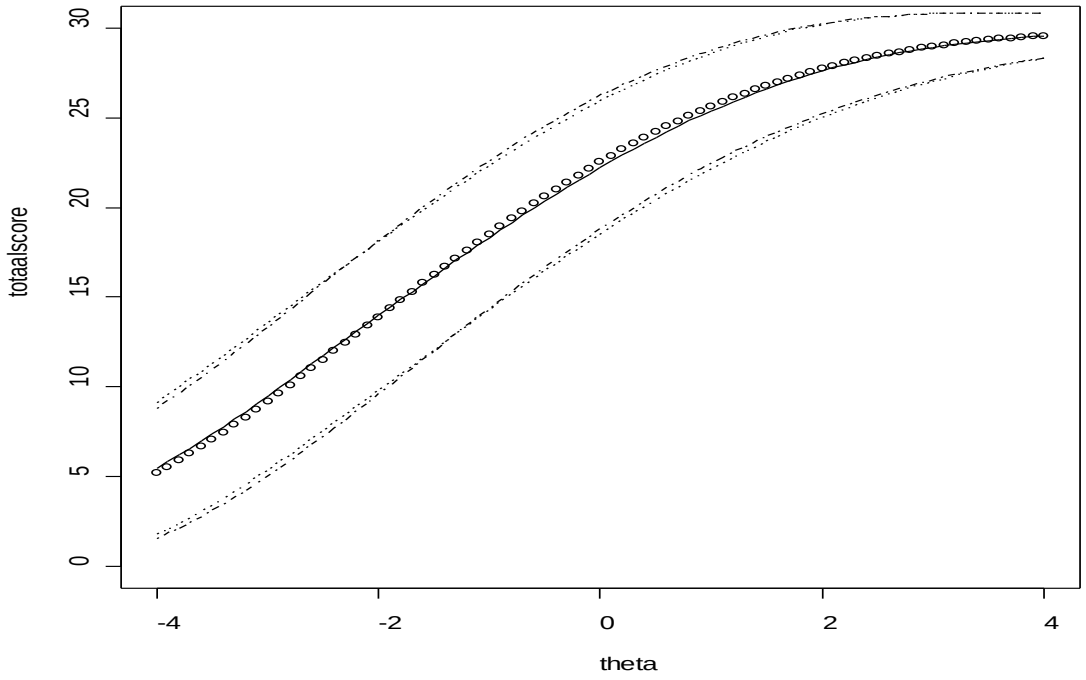
Figuur 4.6 IRFs van Vlamingen (-) en allochtonen (. .) voor items die significante DIF vertonen ($p<.05$)

4.3.2 Verklaren van DIF

Aangezien er maar twee items significante DIF vertonen en het moeilijk is om een cognitieve analyse te doen op deze test die uitsluitend bestaat uit figurale items, zullen we de DIF in de moeilijkheidsgraden niet trachten te verklaren.

4.3.3 Effect van DIF op de testscores

Om het belang van DIF op de testscores na te gaan, worden in Figuur 4.7 per groep de verwachte somscores (en bijhorend betrouwbaarheidsinterval) in functie van θ weergegeven. De verwachte testscore-curves verschillen bijna niet voor beide groepen. De bijhorende betrouwbaarheidsintervallen laten zien dat voor geen enkele waarde van θ er een significant verschil is tussen de verwachte somscores van Vlamingen en allochtonen.



Figuur 4.7 Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

4.3.4 Conclusie

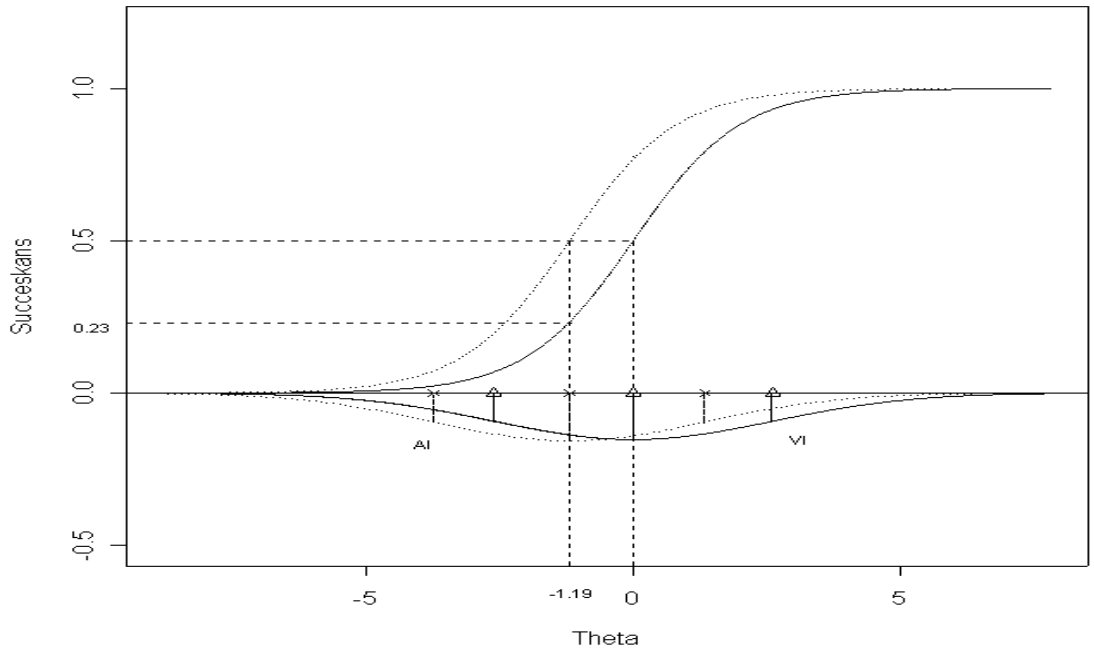
We stellen vast dat de gemiddelde Vlaming beter presteert op deze test dan de gemiddelde allochtoon (μ=-.73, p<.01). Voor een unbiased item met moeilijkheidsgraad gelijk aan 0 heeft de gemiddelde Vlaming (θ = 0) een succeskans van 0.50, terwijl de gemiddelde allochtoon (θ = -.73) een succeskans heeft van 0.32. Verder blijkt dat er slechts in twee items DIF optreedt en dat de praktische significantie van de DIF zeer beperkt is. De mediaan van de absolute verschillen tussen IRFs van Vlamingen en allochtonen is gemiddeld .02 en het 97.5 percentiel van de absolute verschillen tussen IRFs is slechts in twee gevallen groter dan .15. Tenslotte stellen we vast dat de gezamenlijke invloed van DIF in individuele items geen differentieel effect heeft op de verwachte testscores van Vlamingen en allochtonen.

4.4 REKENVAARDIGHEID

4.4.1 Modelleren van DIF

Na schatting van het Raschmodel op elke groep, worden de parameters op één schaal geplaatst met de methode van gelijke populatiegemiddelden.

Voor Vlamingen geldt dat $\theta \sim N(0,2.6^2)$ en voor allochtonen $\theta \sim N(-1.19,2.5^2)$. We stellen dus vast dat in de verzamelde steekproeven Vlamingen gemiddeld beter presteren dan allochtonen
(μ=-1.19, p<.01). Ter illustratie toont Figuur 4.8 de vaardigheidsverdeling voor allochtonen en Vlamingen en de itemresponsfuncties van unbiased items met gemiddelde moeilijkheid voor Vlamingen en allochtonen. We kunnen bijvoorbeeld uit de figuur aflezen dat de gemiddelde Vlaming een succeskans van .50 heeft voor een item met β=0 en dat de succeskans voor een gemiddelde allochtoon op dit item twee keer kleiner is, namelijk .23.



Figuur 4.8 Verdeling van de latente variabele voor Vlamingen (__) en allochtonen (…) en IRFs van unbiased items met gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

Tabel 4.4 geeft een overzicht van de geschatte β-parameters bij Vlamingen en van de geschatte DIF in de moeilijkheidsgraden (ξ) voor allochtonen versus Vlamingen. De laatste twee kolommen van de tabel geven per item informatie over de praktische significantie van de DIF aan de hand van de mediaan en het 95% betrouwbaarheidsinterval van de absolute verschillen tussen de IRFs van beide groepen.

Tabel 4.4 Parameters van de DIF analyse. Mediaan en 95% Betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	SD(β)	ξ	SD(ξ)	Mediaan	95%BI
1	-5.63	.53	-.29	.64	.00	[.00,.03]
2	-4.79	.41	-.41	.50	.00	[.00,.07]
3	-5.42	.50	-.18	.58	.00	[.00,.02]
4	-5.23	.45	-.69	.57	.00	[.00,.09]
5	-3.99	.33	-1.69**	.44	.01	[.00,.32]
6	-3.77	.31	-.74	.40	.01	[.00,.18]
7	-4.44	.36	-.53	.46	.00	[.00,.11]
8	-5.39	.49	.46	.56	.00	[.00,.07]
9	-2.92	.28	-.66	.36	.02	[.00,.16]
10	-5.22	.43	.53	.52	.00	[.00,.10]
11	-2.99	.28	.18	.36	.01	[.00,.04]
12	-3.57	.31	.21	.39	.01	[.00,.05]
13	-3.23	.28	.30	.38	.01	[.00,.08]
14	-3.79	.31	.57	.40	.02	[.00,.14]
15	-2.97	.28	.41	.37	.02	[.00,.10]
16	-2.04	.27	.26	.36	.03	[.00,.06]
17	-1.98	.25	.5	.35	.05	[.00,.12]
18	-1.48	.24	.55	.33	.06	[.00,.14]
19	-.78	.24	.35	.35	.04	[.00,.09]
20	-.52	.24	.40	.35	.04	[.01,.10]
21	-.27	.25	.85**	.35	.09	[.02,.21]
22	-.07	.25	.26	.35	.03	[.01,.06]
23	.37	.25	.37	.37	.04	[.00,.09]
24	.44	.25	.45	.36	.05	[.01,.11]
25	.75	.26	.42	.37	.04	[.00,.10]
26	1.52	.26	.15	.38	.02	[.00,.04]
27	2.01	.28	.04	.39	.00	[.00,.01]
28	2.38	.29	-.54	.41	.05	[.00,.13]
29	2.41	.27	-.72	.39	.07	[.00,.18]
30	2.51	.28	-.77	.41	.07	[.00,.19]

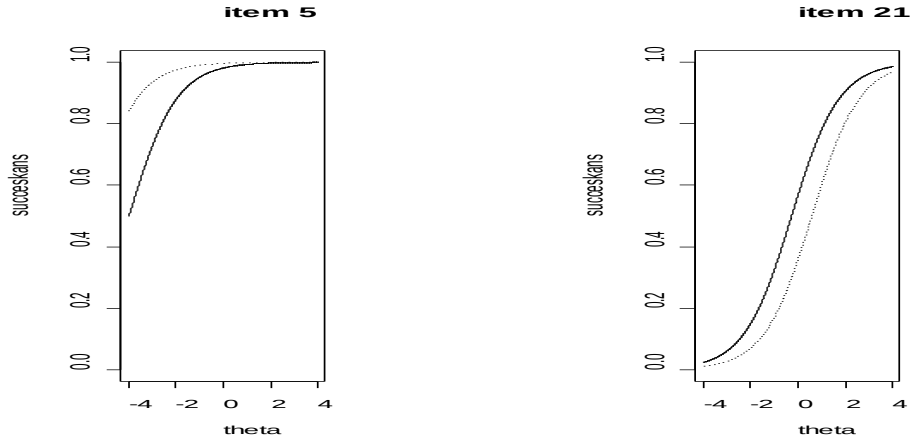
* p<.05; **p<.01

Uit tabel 4.4 kunnen we besluiten dat er bij 2 items (van de 30) significante DIF in de moeilijkheidsgraden optreedt. Item 5 vertoont DIF in het voordeel van de allochtonen en item 21 vertoont DIF in het voordeel van de Vlamingen. De mediaan van de absolute verschillen tussen IRFs van Vlamingen en allochtonen bij items met DIF varieert van 0

tot .09 en heeft een gemiddelde waarde van .03. Het 97.5 percentiel van de absolute verschillen varieert tussen .01 en .32 en heeft een gemiddelde waarde van .11. Absolute verschillen in succeskansen zijn in het algemeen klein, maar bij sommige items en voor een beperkt deel van de schaal zijn de verschillen niet verwaarloosbaar.

Merk op dat de items in de eerste helft van de test weinig informatie geven over de vaardigheid van personen omdat ongeveer iedereen ze juist oplost. Dit komt omdat de test vooral een snelheidstest is : Verschillen in testscores weerspiegelen vooral verschillen in snelheid.

Ter illustratie toont figuur 4.9 de 2 items waar het 97.5 percentiel van de absolute verschillen groter is dan .20. Dit zijn ook de items die significante DIF vertonen.



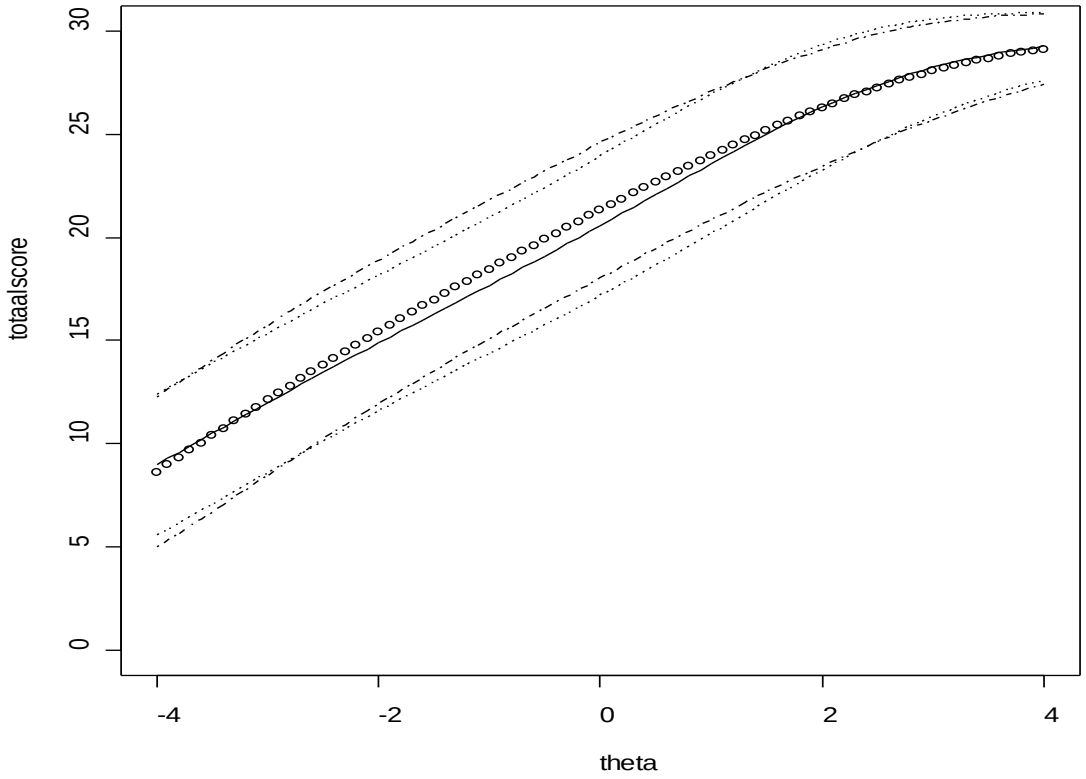
Figuur 4.9 IRFs van Vlamingen (-) en allochtonen (. .) voor items die significante DIF vertonen (p<.01)

4.4.2 Verklaren van DIF

Aangezien er maar twee items significante DIF vertonen, zoeken we geen verklaring voor DIF in de moeilijkheidsgraden.

4.4.3 Effect van DIF op de testscores

Figuur 4.10 toont dat er een klein verschil is in verwachte somscore voor Vlamingen en allochtonen met een θ tussen -2 en 2. Dit verschil (in het voordeel van de Vlamingen) is echter niet significant.



Figuur 4.10 Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

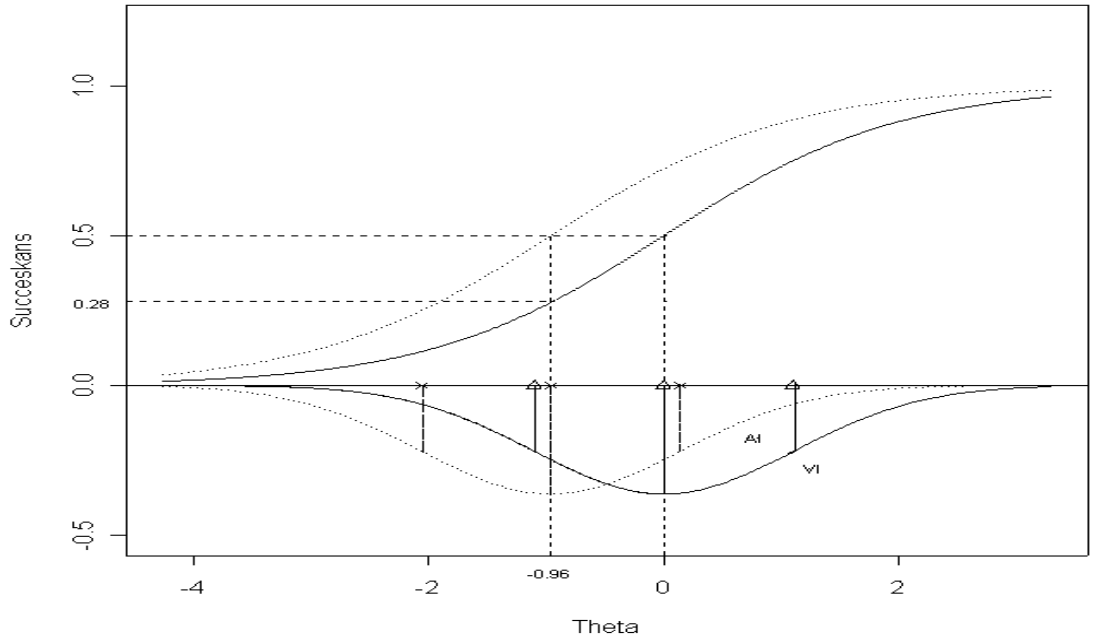
4.4.4 Conclusie

We stellen vast dat de gemiddelde Vlaming in de geobserveerde steekproef gemiddeld beter presteert op de test dan de gemiddelde allochtoon (μ=-1.19, $p<.01$). Een gemiddelde Vlaming heeft ongeveer dubbel zoveel kans dan een gemiddelde allochtoon om een unbiased item met $\beta = 0$ juist op te lossen. Verder blijkt dat item 5 en 21 statistisch significante DIF vertonen (<.01) in de moeilijkheidsgraden. De praktische significantie van de DIF is over het algemeen beperkt, maar is voor enkele items op bepaalde stukken van de schaal wel van betekenis. De DIF in de individuele items heeft een gering effect op de verwachte somscores van Vlamingen en allochtonen met een θ tussen -2 en 2. Dit verschil in verwachte somscores is echter niet significant

4.5 EXCLUSIE

4.5.1 Modelleren van DIF

Het Raschmodel wordt per groep geschat en vervolgens worden de parameters op één schaal geplaatst met de methode van gelijke populatie gemiddelden. Voor Vlamingen geldt dat $\theta \sim N(0,1.1^2)$ en voor allochtonen $\theta \sim N(-0.96,1.1^2)$. We stellen dus vast dat in de verzamelde steekproeven Vlamingen gemiddeld beter presteren dan allochtonen (μ=-.96, p<.01). Ter illustratie toont Figuur 4.11 de vaardigheidsverdeling voor allochtonen en Vlamingen en de itemresponsfuncties van unbiased items met gemiddelde moeilijkheid voor Vlamingen en allochtonen. We kunnen bijvoorbeeld uit de figuur aflezen dat de gemiddelde Vlaming een succeskans van .50 heeft voor een item met β=0 en dat de succeskans voor een gemiddelde allochtoon op dit item twee keer kleiner is, namelijk .28. We kunnen ook aflezen in de Figuur dat de 25% beste allochtonen het net iets beter doen dan de gemiddelde Vlaming.



Figuur 4.11 Verdeling van de latente variabele voor Vlamingen (__) en allochtonen (…) en IRFs van unbiased items met een gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

Tabel 4.5 toont de geschatte itemparameters voor Vlamingen (β) en de geschatte DIF-parameters in de moeilijkheidsgraden (ξ). De ξ-parameters geven aan hoeveel moeilijker een item is voor allochtonen dan voor Vlamingen De laatste twee kolommen van de tabel geven informatie over de praktische significantie van de DIF, namelijk, de mediaan en het 95% betrouwbaarheidsinterval van de absolute verschillen tussen IRFs voor Vlamingen en allochtonen.

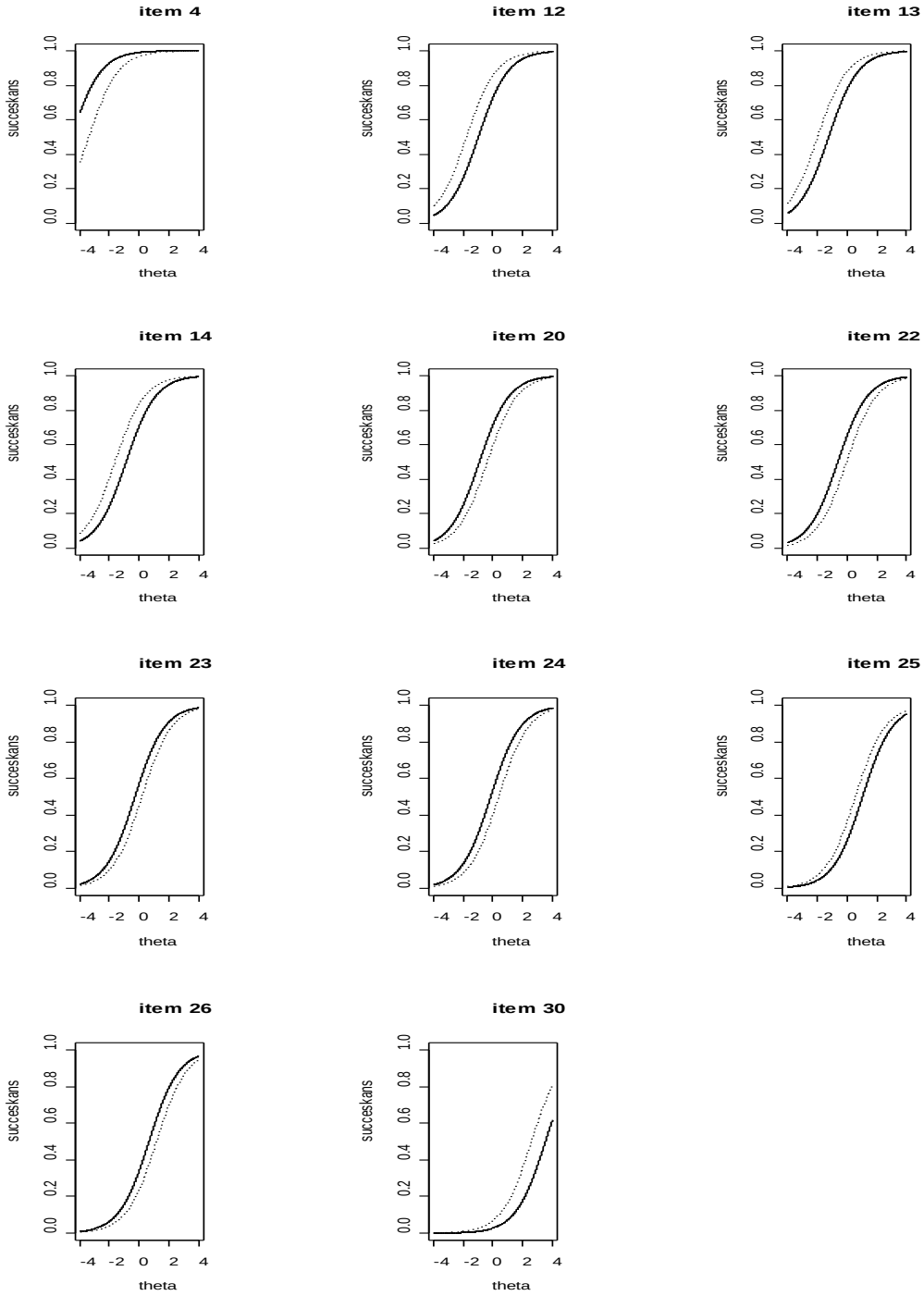
Tabel 4.5 Parameters van de DIF analyse voor de subtest Exclusie. Mediaan en 95% Betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	SD(β)	ξ	SD(ξ)	Mediaan	95% BI
1	-4.02	.38	-.78	.49	.01	[.00,.18]
2	-4.59	.51	.45	.55	.01	[.00,.10]
3	-4.03	.41	.23	.47	.00	[.00,.06]
4	-4.58	.49	1.18*	.52	.02	[.00,.29]
5	-4.35	.46	.80	.50	.02	[.00,.20]
6	-3.18	.29	.49	.34	.02	[.00,.12]
7	-2.83	.26	-.21	.32	.01	[.00,.05]
8	-3.24	.30	-.03	.36	.00	[.00,.01]
9	-2.61	.25	-.43	.32	.02	[.00,.11]
10	-3.65	.32	-.15	.41	.00	[.00,.04]
11	-2.61	.25	.05	.30	.00	[.00,.01]
12	-1.00	.17	-.80**	.23	.09	[.00,.20]
13	-1.25	.18	-.74**	.24	.08	[.00,.18]
14	-.89	.17	-.74**	.23	.08	[.00,.18]
15	-1.28	.18	.11	.23	.01	[.00,.03]
16	-2.49	.24	.06	.28	.00	[.00,.01]
17	-1.25	.18	-.12	.23	.01	[.00,.03]
18	-1.07	.17	-.39	.23	.04	[.00,.10]
19	-1.37	.18	.46	.23	.05	[.00,.11]
20	-.93	.17	.54*	.23	.06	[.00,.13]
21	-.90	.17	-.08	.22	.01	[.00,.02]
22	-.63	.16	.60**	.23	.06	[.00,.15]
23	-.29	.16	.47*	.23	.05	[.00,.12]
24	-.17	.16	.55*	.23	.06	[.00,.14]
25	1.04	.17	-.48*	.23	.05	[.00,.12]
26	.66	.16	.52*	.25	.05	[.00,.13]
27	1.54	.19	-.36	.27	.04	[.00,.09]
28	2.74	.25	.02	.40	.00	[.00,.00]
29	1.45	.18	-.27	.26	.03	[.00,.07]
30	3.52	.32	-.94*	.42	.04	[.00,.23]

* p<.05; **p<.01

Uit Tabel 4.5 blijkt dat de items 4, 20, 22, 23, 24 en 26 significant gemakkelijker zijn voor Vlamingen en dat de items 12, 13, 14, 25 en 30 significant gemakkelijker zijn voor allochtonen. Uit de verdeling van de absolute verschillen tussen IRFs van Vlamingen en allochtonen in Tabel 4.5 blijkt verder dat de praktische significantie van de DIF voor deze

test eerder beperkt is: De mediaan van de absolute verschillen varieert van .00 tot .09 en het 97.5 percentiel van de absolute verschillen is enkel voor item 4 en 30 groter dan .20. De IRFs van items met significante DIF (<.05) worden ter illustratie weergegeven in Figuur 4.12.



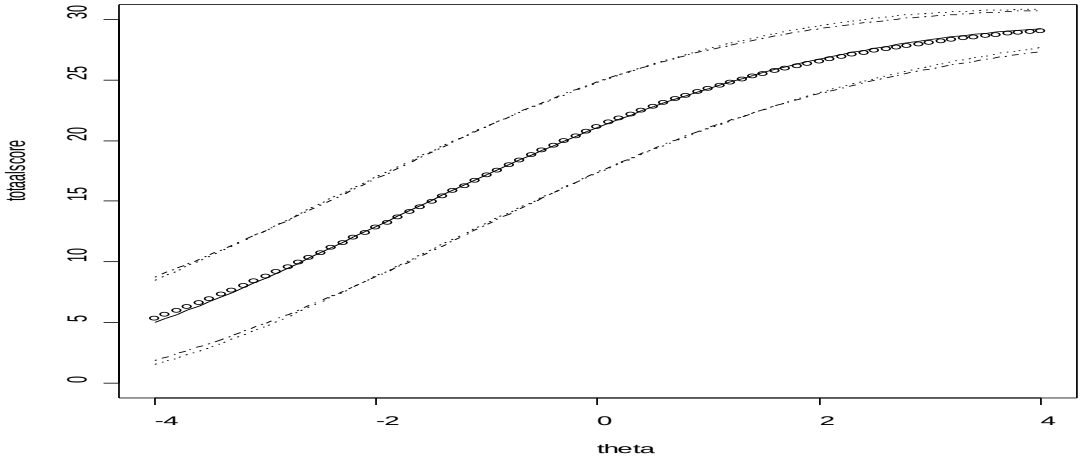
Figuur 4.12 IRFs van Vlamingen (-) en allochtonen (. .) voor items die significante DIF vertonen (p<.05)

4.5.2 Verklaren van DIF

Na het bestuderen van de items werd besloten dat het verklaren van DIF op basis van features heel moeilijk is bij deze items die enkel uit figuraal materiaal bestaan. We beperken ons dan ook tot het verklaren van DIF bij de subtests waar veel DIF aanwezig is en waar de items uit numerieke of verbale stimuli bestaan.

4.5.3 Effect van DIF op de testscores

Om het belang van DIF op de testscores na te gaan, wordt in Figuur 4.13 per groep de verwachte somscores en bijbehorend betrouwbaarheidsinterval in functie van θ weergegeven. De verwachte-somscore curves verschillen bijna niet voor beide groepen. De bijhorende betrouwbaarheidsintervallen laten zien dat voor geen enkele waarde van θ er een significant verschil is tussen de verwachte somscores van Vlamingen en allochtonen. De DIF in individuele testitems heeft dus geen differentieel effect op de (verwachte) somscores en gecorrigeerde somscores van Vlamingen en allochtonen.



Figuur 4.13 Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

4.5.4 Conclusie

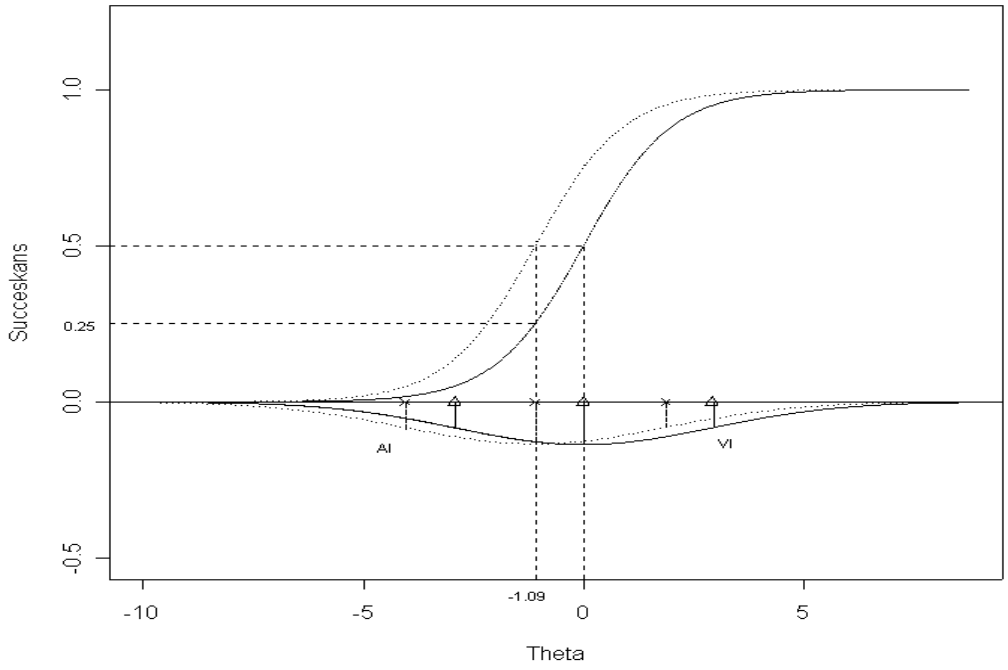
We stellen vast dat de gemiddelde Vlaming beter presteert op de test dan de gemiddelde allochtoon (μ=-.96, p<.01) . Een gemiddelde Vlaming heeft een succeskans van 0.50 voor een unbiased item met β=0 terwijl een gemiddelde allochtoon (θ = -.96) een succeskans van 0.28 heeft. Verder kunnen we stellen dat er significante DIF optreedt in 11 van de 30 items. In 6 items is er DIF in het voordeel van de Vlamingen, in de overige 5 items is er DIF in het voordeel van Allochtonen. De praktische significantie van de DIF is beperkt. De mediaan van de absolute verschillen tussen IRFs van Vlamingen en allochtonen is gemiddeld .03 en het 97.5 percentiel van de absolute verschillen tussen IRFs is slechts in zeven items groter dan .15. Tot slot stellen we vast dat de DIF in individuele items geen differentieel effect heeft op (verwachte) somscores van Vlamingen en allochtonen.

4.6 KONTROLEREN

4.6.1 Modelleren van DIF

Per groep worden de parameters van het Rasch model geschat en op dezelfde schaal geplaatst met de methode van gelijke populatiegemiddelden. Omdat de subtest kontroleren een pure snelheidstest is, is het belangrijk dat het hoofdeffect een goede schatting is van de mate waarin Vlamingen en allochtonen verschillen in snelheid. Als we de methode van gelijke populatiegemiddelden toepassen op alle 100 items van de test, dan wordt het snelheidseffect onderschat omdat dit effect niet zichtbaar is in items die door beide groepen juist werden opgelost of in items die door beide groepen niet bereikt werden. Om een accurate schatting te maken van het hoofdeffect nemen we daarom enkel de items in het midden van de test (items 40-75) in rekening.

Voor Vlamingen geldt dat $\theta \sim N(0,2.9^2)$ en voor allochtonen $\theta \sim N(-1.09,3.0^2)$. We stellen dus vast dat in de verzamelde steekproeven Vlamingen gemiddeld beter presteren dan allochtonen (μ=-1.09, p<.01). Ter illustratie toont Figuur 4.14 de vaardigheidsverdeling voor allochtonen en Vlamingen en de itemresponsfuncties van unbiased items met gemiddelde moeilijkheid voor Vlamingen en allochtonen. We kunnen bijvoorbeeld uit de figuur aflezen dat de gemiddelde Vlaming een succeskans van .50 heeft voor een item met β=0 en dat de succeskans van een gemiddelde allochtoon voor dit item gelijk is aan .25.



Figuur 4.14 Verdeling van de latente variabele voor Vlamingen (__) en allochtonen (…) en IRFs van unbiased items met gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

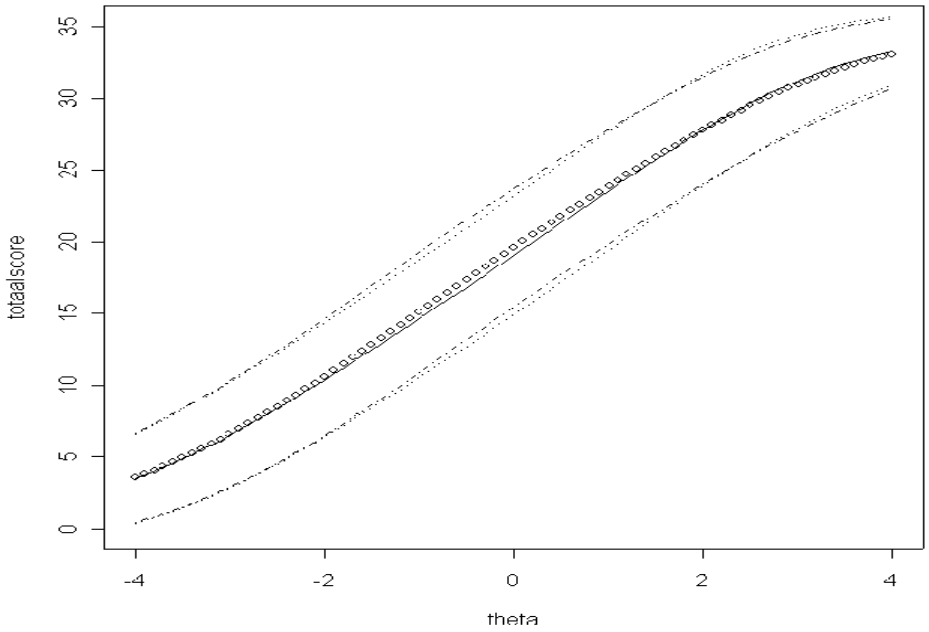
Tabel 4.6 toont de geschatte β-parameters bij Vlamingen en de DIF in de moeilijkheidsgraden (ξ). De twee laatste kolommen van de Tabel bevatten informatie over de praktische significantie van de DIF, meerbepaald, de mediaan en het 95% betrouwbaarheidsinterval van de verdeling van de absolute verschillen tussen de IRFs van Vlamingen en allochtonen.

4.6.2 Verklaren van DIF

Een verklaring van DIF in functie van itemkenmerken heeft bij de subtest "kontroleren" weinig zin omdat deze test een typische snelheidstest is. Alle items zijn in principe even gemakkelijk en als men voldoende tijd heeft kan men in principe elk item juist oplossen.

4.6.3 *Effect van DIF op de testscores*

Om het belang van DIF op de testscores na te gaan, wordt in Figuur 4.16 de verwachte somscores per groep (en bijhorend betrouwbaarheidsinterval) in functie van θ weergegeven. De verwachte somscore-curves verschillen niet significant voor beide groepen, al liggen de verwachte somscores voor Vlamingen met een laag tot gemiddeld vaardigheidsniveau iets hoger dan voor allochtonen met dezelfde vaardigheid (gemiddeld 2 punten).



Figuur 4.16 Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

Tabel 4.6 Parameters van de DIF analyse. Mediaan en 95% BI van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	Sd β	ξ	Sd ξ	Mediaan	95% BI
40	-4.18	.33	.16	.40	.04	[.01,.10]
41	-3.81	.31	.25	.38	.04	[.00,.09]
42	-3.38	.27	-.01	.35	.04	[.00,.09]
43	-3.45	.28	.03	.35	.04	[.00,.09]
44	-3.27	.27	-.07	.36	.04	[.00,.09]
45	-2.17	.25	-.30	.34	.04	[.00,.09]
46	-3.10	.27	.16	.34	.04	[.00,.09]
47	-2.81	.25	-.11	.33	.03	[.00,.08]
48	-1.50	.24	-.58	.33	.03	[.00,.08]
49	-2.22	.24	-.17	.33	.03	[.00,.08]
50	-1.86	.24	.17	.32	.03	[.00,.08]
51	-1.76	.25	-.11	.33	.03	[.00,.08]
52	-1.49	.24	-.01	.31	.03	[.00,.08]
53	-1.46	.24	.17	.32	.03	[.00,.08]
54	-1.00	.24	.18	.34	.03	[.00,.08]
55	-1.07	.24	.24	.33	.03	[.00,.08]
56	-.86	.24	.36	.32	.03	[.00,.08]
57	-.65	.24	.27	.32	.03	[.00,.08]
58	-.55	.24	.44	.33	.03	[.00,.08]
59	-.37	.23	.49	.33	.04	[.00,.08]
60	-.25	.24	.40	.33	.04	[.01,.09]
61	.87	.24	1.23**	.36	.04	[.00,.08]
62	.65	.24	.03	.34	.04	[.00,.08]
63	.91	.25	.14	.35	.04	[.00,.09]
64	1.07	.25	.07	.37	.04	[.01,.09]
65	1.43	.25	-.17	.36	.04	[.01,.09]
66	1.45	.25	.04	.36	.04	[.01,.09]
67	1.63	.26	.21	.37	.04	[.01,.09]
68	1.94	.26	-.40	.38	.04	[.01,.09]
69	2.20	.27	-.50	.38	.04	[.01,.09]
70	2.35	.27	.37	.40	.04	[.01,.10]
71	2.50	.28	-.16	.40	.04	[.01,.10]
72	2.85	.28	-.63	.40	.04	[.01,.10]
73	3.02	.29	-.55	.42	.04	[.01,.10]
74	3.23	.30	-.58	.40	.04	[.01,.10]
75	3.67	.31	-1.03*	.44	.05	[.01,.11]

* p<.05; **p<.01

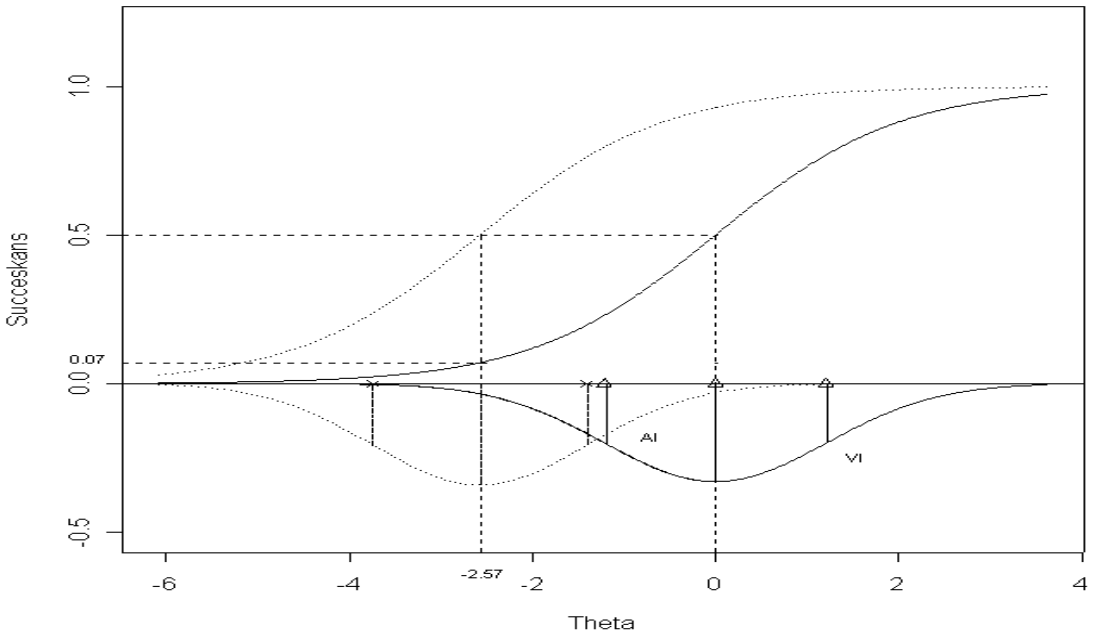
4.6.4 Conclusie

We kunnen besluiten dat allochtonen gemiddeld minder goed presteren op deze test dan Vlamingen (μ=-1.09, p<.01) wat wil zeggen dat ze minder snel werken. Op enkele uitzonderingen na is er geen DIF. De verwachte somscores zijn ook dezelfde in elke groep

4.7 WOORDRELATIES

4.7.1 Modelleren van DIF

De parameters van het Rasch model worden per groep geschat en op dezelfde schaal geplaatst met de methode van gelijke populatiegemiddelden.Voor Vlamingen geldt dat $\theta \sim N(0,1.2^2)$ en voor allochtonen $\theta \sim N(-2.57,1.2^2)$. We stellen dus vast dat in de verzamelde steekproeven Vlamingen gemiddeld beter presteren dan allochtonen (μ=-2.57, p<.01). Dit verschil in gemiddeld presteren is aanzienlijk want het bedraagt 2SDs. Ter illustratie toont Figuur 4.17 de vaardigheidsverdeling voor allochtonen en Vlamingen en de itemresponsfuncties van unbiased items met gemiddelde moeilijkheid voor Vlamingen en allochtonen. We kunnen bijvoorbeeld uit de figuur aflezen dat de gemiddelde Vlaming een succeskans van .50 heeft voor een item met β=0 en dat de succeskans van een gemiddelde allochtoon voor dit item 7 keer kleiner is, namelijk .07. We zien ook dat de 25% laagst scorende Vlamingen (percentiel 25 in distributie θ voor Vlamingen) nog beter scoren dan de 25% hoogst scorende allochtonen (percentiel 75 in distributie θ voor allochtonen).



Figuur 4.17 Verdeling van de latente variabele voor Vlamingen (__) en allochtonen (…) en IRFs van unbiased items met gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

Tabel 4.7 geeft een overzicht van de geschatte itemparameters voor de Vlamingen (β) en van het verschil in moeilijkheidsgraden (ξ) voor Vlamingen en allochtonen en beschrijft de praktische significantie van de DIF aan de hand van de mediaan en het 95% BI van de absolute verschillen tussen de IRFs van Vlamingen en allochtonen.

Tabel 4.7 Parameters van de DIF analyse voor de subtest Woordrelaties. Mediaan en 95% Betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	SD(β)	ξ	SD(ξ)	Mediaan	95% BI
1	-4.61	.48	-1.05	.54	.01	[.00,.17]
2	-4.39	.42	-1.59**	.49	.01	[.00,.25]
3	-4.86	.56	.77	.58	.01	[.00,.17]
4	-1.85	.20	-2.15**	.26	.12	[.00,.49]
5	-2.95	.25	-1.30**	.31	.04	[.00,.31]
6	-2.66	.24	.04	.27	.00	[.00,.01]
7	-3.79	.37	1.55**	.39	.07	[.00,.37]
8	-4.41	.47	.65	.50	.01	[.00,.16]
9	-2.32	.22	1.01**	.28	.11	[.00,.25]
10	-.82	.17	-1.33**	.22	.14	[.01,.32]
11	-1.63	.18	.18	.24	.02	[.00,.04]
12	-1.72	.18	.40	.24	.04	[.00,.10]
13	-2.30	.22	1.17**	.28	.13	[.00,.28]
14	-1.65	.18	.43	.25	.05	[.00,.11]
15	-1.32	.18	.19	.24	.02	[.00,.05]
16	-.47	.17	-1.11**	.24	.12	[.01,.27]
17	-2.21	.22	-.26	.26	.02	[.00,.06]
18	-2.17	.21	-.03	.26	.00	[.00,.01]
19	-.08	.16	-.60*	.24	.06	[.01,.15]
20	-1.15	.17	-.81**	.23	.09	[.00,.20]
21	-1.36	.18	.43	.25	.05	[.00,.11]
22	-1.63	.19	.75**	.27	.08	[.00,.19]
23	-.91	.17	1.21**	.29	.13	[.02,.29]
24	-1.50	.18	1.60**	.29	.17	[.02,.38]
25	-.68	.17	-.22	.25	.02	[.00,.06]
26	-3.01	.26	2.37**	.32	.27	[.01,.53]
27	-1.71	.19	1.63**	.29	.18	[.01,.39]
28	-.39	.16	-.09	.26	.01	[.00,.02]
29	-1.40	.18	1.36**	.28	.15	[.02,.33]
30	-1.10	.17	.94**	.28	.10	[.01,.23]
31	.00	.16	-.69**	.25	.07	[.01,.17]
32	.81	.17	-.35	.31	.04	[.00,.09]
33	-.26	.17	.02	.27	.00	[.00,.01]
34	1.23	.17	-1.23**	.27	.13	[.02,.30]
35	2.46	.22	-1.13**	.40	.12	[.00,.28]
36	1.02	.17	-.56	.30	.06	[.01,.14]
37	1.04	.17	-.85**	.30	.09	[.01,.21]
38	2.06	.21	-1.91**	.31	.21	[.01,.44]
39	.84	.18	.40	.37	.04	[.00,.10]
40	2.00	.20	.26	.51	.02	[.00,.06]
41	2.56	.23	-1.24**	.41	.13	[.00,.30]
42	.74	.18	1.33**	.50	.14	[.01,.32]
43	1.55	.19	-1.50**	.30	.16	[.02,.36]
44	1.46	.19	.26	.44	.03	[.00,.06]
45	1.54	.19	1.04	.59	.11	[.00,.25]

* p<.05; **p<.01

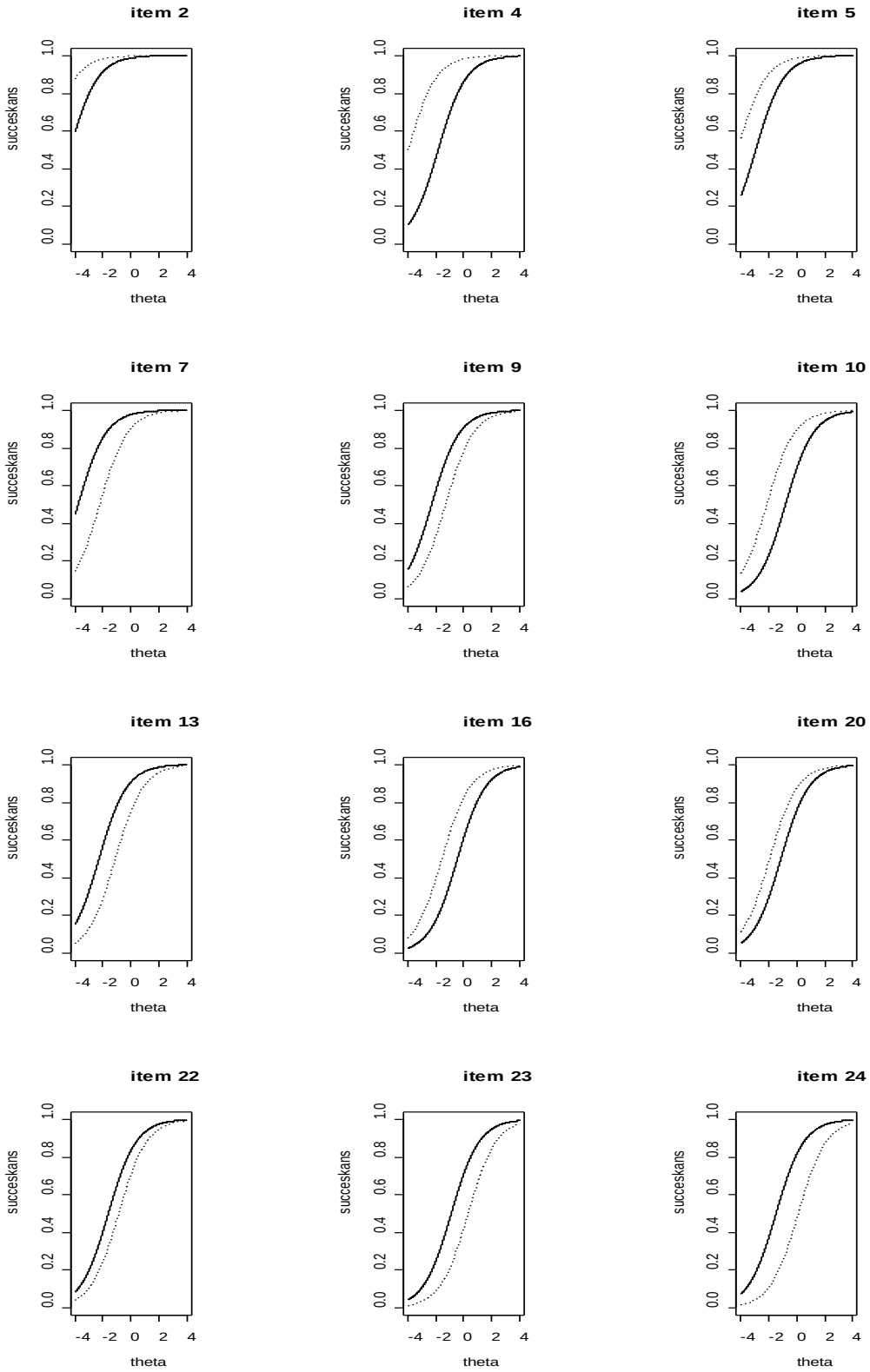
Uit Tabel 4.7 blijkt dat 25 van de 45 items uniforme DIF vertonen. Hiervan zijn er 12 in het voordeel van de allochtonen en 13 in het voordeel van de Vlamingen. Merk op dat het hoofdeffect zoals bepaald met de methode van gelijke populatiegemiddelden leidt tot zowel positieve als negatieve DIF in items. De globale proportie juist toont dat alle items moeilijker zijn voor allochtonen (zie figuur in sectie 2.7), maar na correctie voor het verschil in gemiddeld presteren zijn sommige items relatief gemakkelijker (negatieve ξ) en zijn andere items relatief moeilijker (positieve ξ) voor allochtonen.

De praktische significantie van de DIF is voor veel items relatief groot. De mediaan van de absolute verschillen varieert van .00 tot .27, met een gemiddelde van .08.
Het 97.5 percentiel van de absolute verschillen in succeskansen varieert van .00 tot .53, met een gemiddelde van .21. Voor 24 items is het 97.5 percentiel groter dan .15, wat betekent dat in ongeveer 53% (24/45) van de items de verschillen tussen succeskansen van Vlamingen en allochtonen op een klein stuk van de latente schaal (2.5 %) minstens .15 zijn. Figuur 4.18 toont de IRFs voor items die significante uniforme DIF vertonen ($p<.01$).

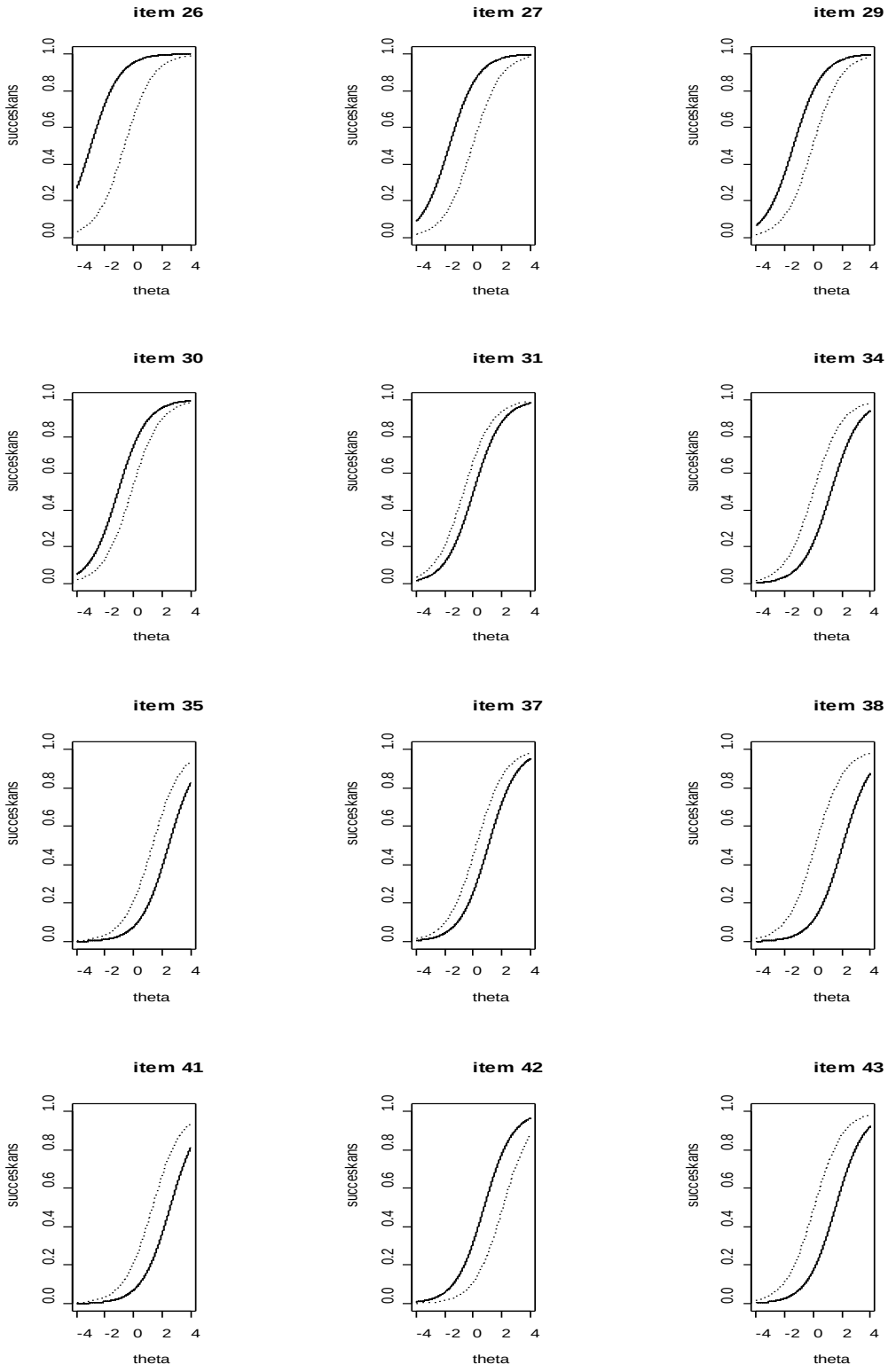
4.7.2 Verklaren van DIF

Om moeilijkheidsgraden en verschillen in moeilijkheidsgraden van items in verbale subtests te verklaren werden 4 verschillende variabelen opgesteld.

- De variabele FREQ meet de gemiddelde woordfrequentie van de 4 woorden die deel uit maken van het item. De woordfrequentie is gebaseerd op een database (CELEX) met woordfrequenties van Nederlandse woorden. We veronderstellen dat woordfrequentie een rol speelt omdat woorden die veel voorkomen in gesproken of geschreven Nederlands waarschijnlijk ook door allochtonen beter gekend zijn.
- De variabele SYNO meet het soort van relatie dat moet gezocht worden in het item (SYNO=1 als het item 2 synoniemen bevat en SYNO=0 als het item 2 antoniemen bevat).
- De variabele ABSTRACT geeft aan hoe abstract itemwoorden zijn (1=heel concreet,...,5=heel abstract). De variabele is berekend als het gemiddeld oordeel van twee NT2 instructeurs (Cronbach's α = .81) over de 4 itemwoorden. We veronderstellen dat deze variabele een rol speelt omdat abstracte woorden waarschijnlijk minder gekend zijn voor allochtonen die meestal een andere moedertaal hebben.
- De variabele VREEMD geeft aan in welke mate men de betekenis van itemwoorden kan afleiden op basis van een vreemde taal (1=heel moeilijk,.., 5=heel gemakkelijk). De variabele is berekend als het gemiddeld oordeel van drie NT2 instructeurs (Cronbach's α = .79) over de 4 itemwoorden. We nemen aan dat deze variabele een rol speelt omdat woorden die verwant zijn aan een vreemde taal (vb tactiek; engels: tactic) gemakkelijker kunnen zijn voor allochtonen.



Figuur 4.18 IRFs van Vlamingen (-) en allochtonen (. .) voor items die significante DIF vertonen (p<.01)



Vervolg Figuur 4.18 IRFs van Vlamingen (-) en allochtonen (. .) voor items die significante DIF vertonen (p<.01)

Regressieanalyse van de geschatte DIF parameters en van de moeilijkheidsgraden per groep op deze 4 verklarende variabelen heeft het volgende resultaat (zie Tabel 4.9):

Items zijn makkelijker voor zowel Vlamingen als Allochtonen wanneer ze woorden bevatten die vaker voorkomen. Verder blijkt dat items met abstracte woorden moeilijker zijn voor zowel allochtonen als voor Vlamingen dan items met concrete woorden. Wanneer het item gemakkelijker kan gekend worden op basis van een vreemde taal dan wordt het moeilijker voor Vlamingen, terwijl dit voor allochtonen geen verschil maakt. Of men nu synoniemen of antoniemen moet zoeken, dit maakt de items voor beide groepen noch moeilijker noch gemakkelijker. Op basis van deze vijf variabelen wordt 45 % van de moeilijkheid van de items bij Vlamingen en 52 % van de moeilijkheid van de items bij allochtonen verklaard.

Tabel 4.9 Gestandaardiseerde regressiegewichten van analyse waarbij de moeilijkheidsgraden per groep en het verschil in moeilijkheidsgraden gemodelleerd worden in functie van de afhankelijke variabelen

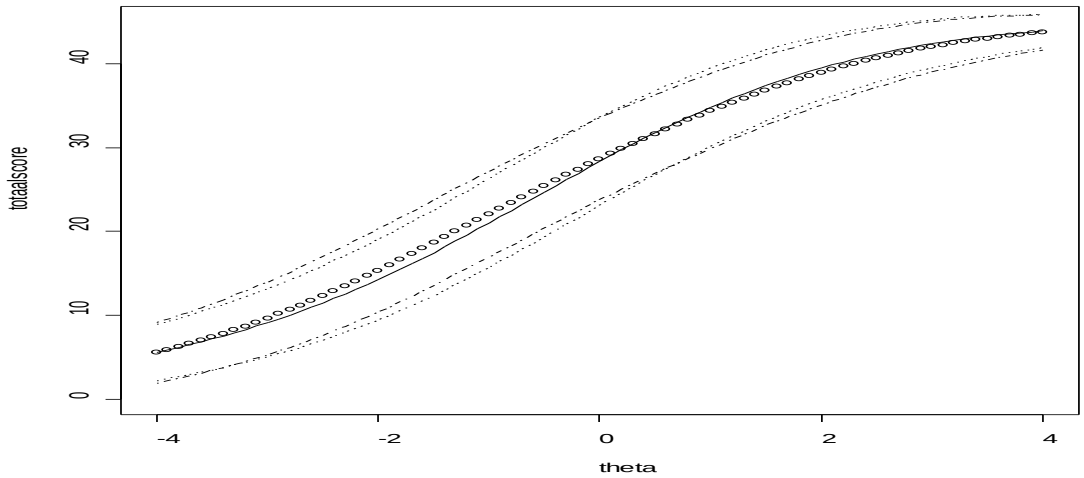
OV	AV	gewicht	p-waarde	R^2
	FREQ	-0,37	0,01	
βvl	SYNO	-0,11	0,35	0.45
	ABSTRACT	0,38	0,009	
	VREEMD	0,25	0,04	
	FREQ	-0,52	0,0002	
βal	SYNO	0,08	0,49	0.52
	ABSTRACT	0,29	0,03	
	VREEMD	-0,06	0,55	
	FREQ	-0,27	0,04	
$\xi = \beta_{al} - \beta_{vl}$	SYNO	0,35	0,005	0.49
	ABSTRACT	-0,17	0,20	
	VREEMD	-0,57	<0,0001	

Wanneer we het verschil in moeilijkheidsgraden trachten te verklaren stellen we het volgende vast: Een item wordt gemakkelijker voor allochtonen (in vergelijking met andere items) als de woorden van het item meer voorkomen in het Nederlands of als ze verwant zijn aan een andere taal en als men op zoek moet gaan naar een tegenstelling eerder dan naar een synoniem. In deze analyse wordt 49% variantie verklaard van het verschil in moeilijkheid tussen Vlamingen en allochtonen.

Met deze bevindingen kan men in de toekomst trachten rekening te houden bij de constructie van verbale intelligentietests. Zodra één of meerdere van deze kenmerken (die in het nadeel spelen van een bepaalde groep) vaak voorkomen in de items van een test, dan verhoogt de kans op itembias.

4.7.3 Effect van DIF op de testscore

Het belang van DIF op de testscores wordt in Figuur 4.19 weergegeven. Deze figuur toont de verwachte somscores (en bijbehorend betrouwbaarheidsinterval) per groep in functie van θ. De somscore-curves verschillen een beetje voor personen met een lage tot gemiddelde vaardigheid, maar voor geen enkele waarde van θ er een significant verschil tussen de verwachte somscores van Vlamingen en allochtonen. De vele DIF-items hebben dus geen differentieel effect op de verwachte somscores.



Figuur 4.19 Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

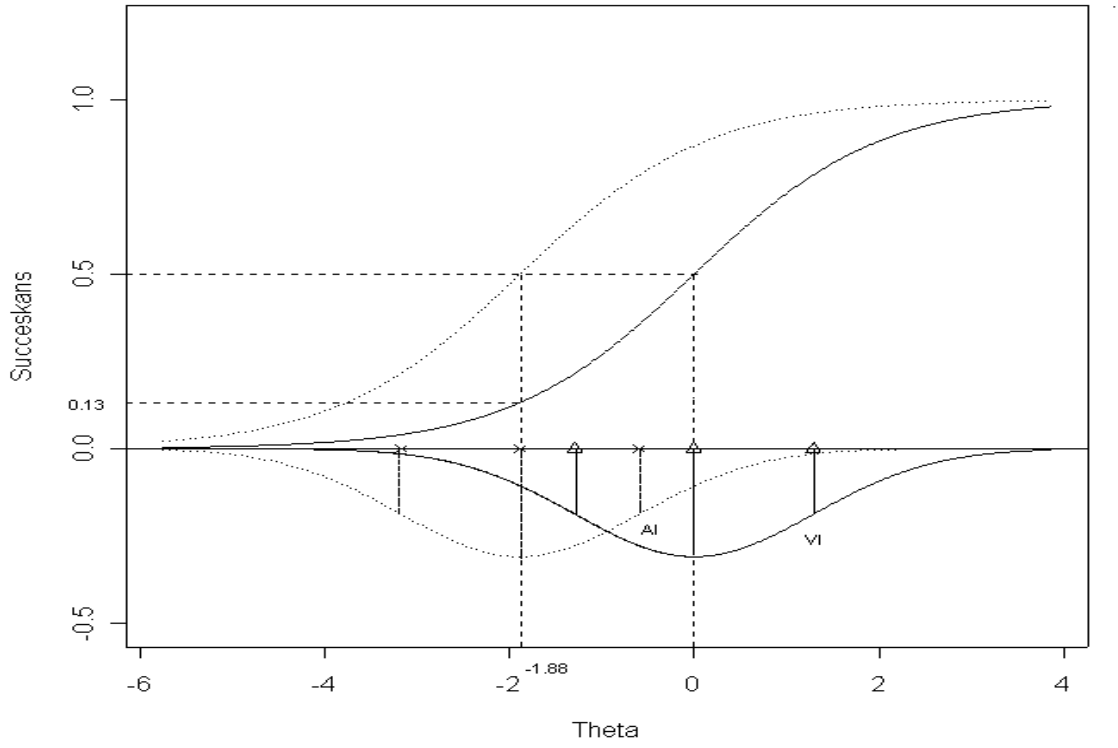
4.7.4 Conclusie

Een opvallende vaststelling is dat Vlamingen gemiddeld veel beter presteren op de test dan allochtonen. Dit verschil tussen beide groepen is groter dan bij de niet-verbale subtests. Meerbepaald blijkt dat de gemiddelde Vlaming 7 keer meer kans heeft om een unbiased item met gemiddelde moeilijkheidsgraad op te lossen dan een gemiddelde allochtoon. Verder blijkt ook dat er DIF optreedt in 25 van de 45 items. Hiervan zijn er 12 in het voordeel van de allochtonen en 13 in het voordeel van de Vlamingen. De praktische significantie van de DIF is vrij groot. De mediaan van de absolute verschillen tussen IRFs van Vlamingen en allochtonen is gemiddeld .08 en het 97.5 percentiel van de absolute verschillen tussen IRFs is in 24 items groter dan .15.
Items die meer concrete woorden bevatten of woorden die frequent voorkomen in de Nederlandse taal blijken significant gemakkelijker te zijn voor zowel allochtonen als Vlamingen. De vastgestelde DIF kan voor 49% verklaard worden op basis van de itemkenmerken: Items zijn relatief gemakkelijker voor allochtonen als de woorden van het item meer voorkomen in het Nederlands of als men de woorden kan kennen op basis van een vreemde taal. Verder blijken items waar men een synoniem moet zoeken voor allochtonen moeilijker dan items waar men een tegenstelling moet zoeken. De DIF in de individuele items leidt niet tot een verschil in verwachte somscores van Vlamingen en allochtonen.

4.8 WOORDANALOGIEËN

4.8.1 Modelleren van DIF

Het Raschmodel wordt per groep geschat en de parameters worden op één schaal geplaatst met de methode van gelijke populatie gemiddelden. Voor Vlamingen geldt dat $\theta \sim N(0,1.3^2)$ en voor allochtonen $\theta \sim N(-1.88,1.3^2)$. We stellen dus vast dat in de verzamelde steekproeven Vlamingen gemiddeld beter presteren dan allochtonen (μ=-1.88, $p<.01$). Ter illustratie toont Figuur 4.20 de vaardigheidsverdeling voor allochtonen en Vlamingen en de itemresponsfuncties van unbiased items met gemiddelde moeilijkheid voor Vlamingen en allochtonen. We kunnen bijvoorbeeld uit de figuur aflezen dat de gemiddelde Vlaming een succeskans van .50 heeft voor een item met β=0 en dat de succeskans van een gemiddelde allochtoon bijna 4 keer kleiner is, namelijk .13.



Figuur 4.20. Verdeling van de latente variabele voor Vlamingen (__) en allochtonen (…) en IRFs van unbiased items met gemiddelde moeilijkheid voor Vlamingen (__) en allochtonen (…).

In Tabel 4.10 worden de geschatte moeilijkheidsgraden van Vlamingen (β) en de DIF-parameters (ξ) weergegeven, evenals de mediaan en het 95% betrouwbaarheidsinterval van de verdeling van de absolute verschillen tussen de IRFs van Vlamingen en allochtonen.

Tabel 4.10 Parameters van de DIF analyse voor de subtest Woordanalogieën. Mediaan en 95% Betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor Vlamingen en allochtonen

Item	β	SD(β)	ξ	SD(ξ)	Mediaan	95% BI
1	-2.18	.22	-1.56**	.29	.08	[.00,.37]
2	-3.49	.31	-.19	.36	.00	[.00,.05]
3	-3.04	.28	.69*	.33	.04	[.00,.17]
4	-1.77	.20	-1.12**	.26	.09	[.00,.27]
5	-3.03	.27	-.32	.33	.01	[.00,.08]
6	-3.59	.33	.51	.37	.02	[.00,.13]
7	-3.26	.29	-.80*	.36	.02	[.00,.20]
8	-3.40	.31	-.15	.36	.00	[.00,.04]
9	-2.68	.25	-.30	.31	.02	[.00,.08]
10	-2.03	.21	-.47	.27	.04	[.00,.12]
11	-2.55	.24	-.72*	.30	.04	[.00,.18]
12	-2.16	.22	-1.18**	.27	.07	[.00,.29]
13	-1.67	.21	-1.38**	.27	.11	[.00,.33]
14	-1.64	.20	-.08	.26	.01	[.00,.02]
15	-3.24	.29	.59	.33	.03	[.00,.15]
16	-2.13	.21	.78**	.27	.08	[.00,.19]
17	-2.42	.23	1.07**	.27	.11	[.00,.26]
18	-2.89	.27	1.17**	.31	.10	[.00,.28]
19	-1.36	.19	-.06	.25	.01	[.00,.02]
20	-3.04	.28	.10	.32	.00	[.00,.03]
21	-1.10	.18	-.89**	.24	.09	[.00,.22]
22	-.74	.18	-.35	.24	.04	[.00,.09]
23	-2.07	.21	.47	.27	.05	[.00,.12]
24	-3.03	.27	2.00**	.31	.22	[.01,.46]
25	-.58	.17	-.88**	.22	.09	[.01,.22]
26	-.30	.17	-.57*	.24	.06	[.01,.14]
27	-1.19	.18	.43	.26	.05	[.00,.11]
28	-2.37	.23	2.44**	.30	.28	[.02,.54]
29	-1.84	.21	.39	.27	.04	[.00,.10]
30	.08	.17	.36	.27	.04	[.01,.09]

* p<.05; **p<.01

Uit Tabel 4.10 blijkt dat de helft van de items significante uniforme DIF vertoont. 9 van deze items vertonen DIF in het voordeel van de allochtonen, de overige 6 items vertonen DIF in het voordeel van Vlamingen. Op basis van de globale proportie juist (zie sectie 2.7) stellen we dus vast alle items moeilijker zijn voor allochtonen dan voor Vlamingen

(=hoofdeffect), maar na correctie voor dit verschil in gemiddeld presteren blijkt dat sommige items relatief gemakkelijker zijn voor allochtonen en andere relatief moeilijker.

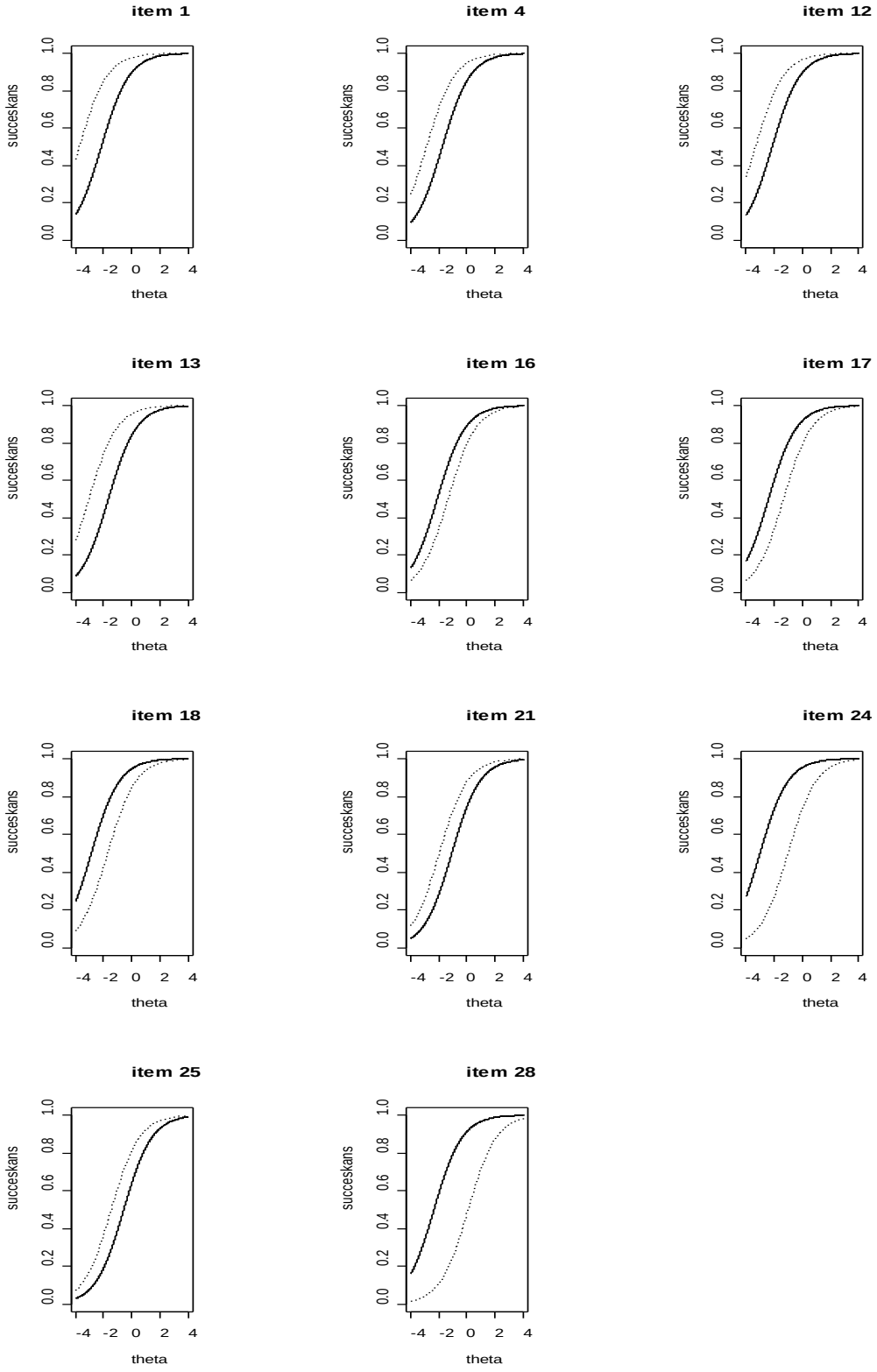
De praktische significantie van de DIF is voor bepaalde items relatief groot. De mediaan van de absolute verschillen varieert van .00 tot .28 en voor 14 items is het 97.5 percentiel groter dan .15. In ongeveer 47% (14/30) van de items zijn de verschillen tussen succeskansen van Vlamingen en allochtonen op een klein stuk van de latente schaal (2.5 %) minstens .15. Figuur 4.21 toont de IRFs voor items die significante uniforme DIF vertonen ($p<.01$).

4.8.2 Verklaren van DIF

Net zoals bij de subtest Woordrelaties trachten we de moeilijkheid van de items voor Vlamingen en allochtonen en het verschil in moeilijkheidsgraden tussen Vlamingen en allochtonen te verklaren aan de hand van enkele itemkenmerken. De volgende variabelen werden opgesteld:

- De variabele ANALOG beschrijft het type van analogie dat moet gezocht worden (1=vergelijking, 2=tegenstelling, 3=specificatie). Om de informatie in deze polytome variabele te coderen worden er twee binaire variabelen gebruikt:
  - de variabele TEGEN is gelijk aan 1 als de analogie een tegenstelling is en 0 anders
  - de variabele DEEL is 1 als de relatie 'A is een deel van B' gezocht moet worden en 0 anders

  Categorie 1 van de variabele ANALOG wordt als basiscategorie gebruikt.
- De variabele FREQ meet de gemiddelde woordfrequentie van de woorden die deel uitmaken van het item (woorden in opgave en antwoordalternatieven). De woordfrequentie is gebaseerd op een database (CELEX) met woordfrequenties van Nederlandse woorden. We veronderstellen dat woordfrequentie een rol speelt omdat woorden die veel voorkomen in gesproken of geschreven Nederlands waarschijnlijk ook door allochtonen beter gekend zijn.
- De variabele ABSTRACT geeft aan hoe abstract itemwoorden zijn (1=heel concreet,...,5=heel abstract). De variabele is berekend als het gemiddeld oordeel van twee NT2 instructeurs over de itemwoorden. We veronderstellen dat deze variabele een rol speelt omdat abstracte woorden waarschijnlijk minder gekend zijn voor allochtonen die meestal een andere moedertaal dan Nederlands hebben.
- De variabele VREEMD geeft aan in welke mate men de betekenis van itemwoorden kan afleiden op basis van een vreemde taal (1=heel moeilijk,.., 5=heel gemakkelijk). De variabele is berekend als het gemiddeld oordeel van drie NT2 instructeurs over de itemwoorden. We nemen aan dat deze variabele een rol speelt omdat woorden die verwant zijn aan een vreemde taal gemakkelijker kunnen zijn voor allochtonen.



Figuur 4.21 IRFs van Vlamingen (-) en allochtonen (. .) voor items die significante DIF vertonen (p<.01)

De resultaten tonen dat alleen de variabelen FREQ en VREEMD een rol spelen. Als we deze variabelen apart opnemen in een analyse geeft dat de resultaten in Tabel 4.11.

Tabel 4.11 Gestandaardiseerde regressiegewichten van analyse waarbij de moeilijkheidsgraden per groep en het verschil in moeilijkheidsgraden gemodelleerd worden in functie van de itemkenmerken.

criterium	predictor	gewicht	p-waarde	R^2
β_{vl}	FREQ	-0,38	0,03	0,25
	VREEMD	-0,29	0,09	
β_{al}	FREQ	-0,08	0,69	0,02
	VREEMD	0,12	0,53	
$\xi = \beta_{al} - \beta_{vl}$	FREQ	-0,39	0,002	0,43
	VREEMD	-0,49	0,01	

We verklaren 43% van de variantie in ξ met behulp van de twee predictoren FREQ en VREEMD. De items van de test Woordanalogieën zijn moeilijker voor allochtonen (in vergelijking met Vlamingen) wanneer de items woorden bevatten die niet frequent voorkomen in het Nederlands of wanneer de woorden minder gemakkelijk kunnen afgeleid worden van een andere taal. De moeilijkheidsgraden van de items kunnen maar zeer beperkt verklaard worden op basis van de itemkenmerken, respectievelijk 25 % variantie van β_{vl} en 2% van β_{al} kan verklaard worden.

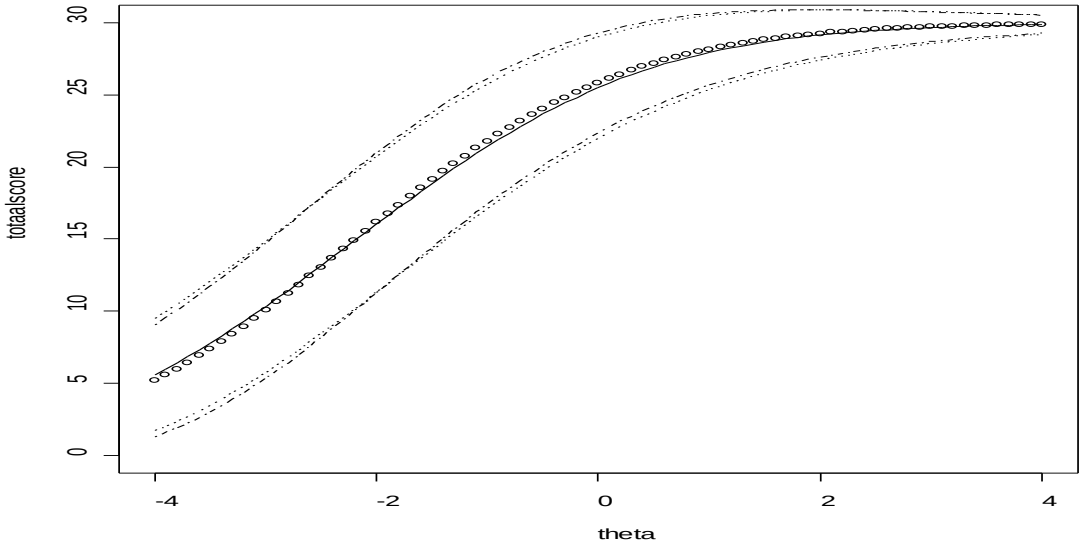
4.8.3 Effect van DIF op de testscore

In Figuur 4.14 worden de verwachte somscores en bijhorend betrouwbaarheidsinterval per groep in functie van θ weergegeven. De verwachte testscore-curves verschillen bijna niet voor beide groepen. Voor geen enkele waarde van θ er een significant verschil tussen de verwachte testscores van Vlamingen en allochtonen.

4.8.4 Conclusie

We stellen vast dat Vlamingen gemiddeld veel beter presteren op deze test met verbale analogieën dan allochtonen (μ=-1.88 p<.01). De gemiddelde Vlaming heeft bijna 4 keer meer kans dan een gemiddelde allochtoon om een unbiased item van gemiddelde moeilijkheid juist op te lossen. Bij de helft van de items treedt uniforme DIF op, die soms ook praktische significant is. De mediaan van de absolute verschillen tussen IRFs van Vlamingen en allochtonen is gemiddeld .06 en voor bijna de helft van de items, is het 97.5 percentiel van de absolute verschillen groter dan .15. Woordanalogieën zijn moeilijker voor allochtonen (in vergelijking met Vlamingen) wanneer de items van de test woorden bevatten die niet frequent voorkomen in de Nederlandse taal en wanneer de woorden typisch Nederlands zijn en niet kunnen afgeleid worden van een andere taal.

De gezamenlijke invloed van DIF in individuele items heft elkaar op, zodat er geen verschillen verwacht worden in de somscores voor Vlamingen en allochtonen



Figuur 4.14 Verwachte testscore voor Vlamingen (o) en allochtonen (-) en 95% betrouwbaarheidsinterval voor Vlamingen (_ . _) en allochtonen (. . .)

Hoofdstuk 5: Achtergrondvariabelen die verband houden met de vaardigheid (θ)

Bij de afname van de MCT-M werd aan elke kandidaat gevraagd om een gegevensformulier in te vullen. Dit formulier (in bijlage **) verschaft de volgende informatie over kandidaat: nationaliteit (nu en bij geboorte), geboorteland, geboortedatum, geslacht, verblijfsduur in België, nationaliteit van ouders en grootouders (bij geboorte), geboorteland van vader/moeder, moedertaal, spreektaal, gevolgde opleidingen, of men een opleiding Nederlands heeft gevolgd en hoe lang, hoeveel jaar betaald werk men heeft gehad in België/Buitenland, ervaring met intelligentietests/persoonlijkheidstests en meerkeuzevragen. Daarnaast werd van elke allochtone kandidaat en van een beperkt aantal Vlamingen de taaltest afgenomen. Op basis van deze informatie hebben we voor beide groepen de volgende variabelen opgesteld:

- VROUW: Deze variabele is gelijk aan 1 voor vrouwen en 0 voor mannen.

- LEEFTIJD: Dit is een continue variabele met de leeftijd van de kandidaat.

- De variabele DIPLOMA beschrijft het hoogst behaalde diploma van de kandidaat (1=basisonderwijs, 2=lager secundair, 3= hoger secundair, 4=hoger niet-universitair, 5=universitair). Deze nominale variabele wordt gecodeerd met 4 binaire variabelen. De categorie hoger secundair wordt als basiscategorie gebruikt. De 4 binaire coderingsvariabelen zijn als volgt gedefinieerd:
  - BASIS=1 als DIPLOMA=1 en anders 0
  - LAGSEC=1 als DIPLOMA=2 en anders 0
  - HOGNTUNIV=1 als DIPLOMA=4 en anders 0
  - UNIV=1 als DIPLOMA=5 en anders 0

- TAALSCORE is een continue variabele die de score van de kandidaat op de MCT-M-Taaltest uitdrukt. Deze score varieert van 40 tot 80.

- TESTERV: Deze variabele is 1 als de kandidaat vertrouwd is met persoonlijkheids- en/of intelligentietests tests en 0 anders.

- MCERV: Deze variabele is 1 als de kandidaat vertrouwd is met meerkeuzevragen en 0 anders.

Specifiek voor allochtonen werden ook nog de onderstaande variabelen opgesteld:

- De variabele IMMIGRATIE beschrijft het tijdstip van immigratie (1=in België geboren, 2=voor 7 jaar naar België geïmmigreerd, 3= na 7 jaar naar België geïmmigreerd. De variabele wordt gecodeerd met 2 binaire variabelen:
  - IMMIV7=1 als geïmmigreerd voor 7 jaar en 0 anders

- IMMIN7=1 als geïmmigreerd na 7 en 0 anders

- De nationaliteit van de ouders wordt gecodeerd met twee binaire variabelen:
  - OUDERSVL=1 als beide ouders Vlaams zijn en anders 0
  - OUDERSAL=1 als beide ouders allochtoon zijn

- MOENED: Deze variabele is 1 als de kandidaat Nederlands als moedertaal heeft en 0 anders.

- SPREENED: Deze variabele is 1 als de kandidaat meestal Nederlands spreekt en 0 anders.

- VERBLIJF: Dit is een continue variabele met het aantal jaren dat de kandidaat in België verblijft.

Deze laatste lijst variabelen zijn alleen zinvol voor allochtonen. De Vlaamse steekproef bevat enkel kandidaten die in België geboren zijn, met twee Vlaamse ouders, en die Nederlands als moedertaal en spreektaal hebben.

Om te onderzoeken in welke mate scores op subtests kunnen verklaard worden op basis van achtergrondvariabelen van kandidaten voeren we per groep (allochtonen, Vlamingen) een multiple regressie uit met bovenstaande achtergrondvariabelen als predictoren. Als afhankelijke variabele in de regressie gebruiken we een minimaal aantal factoren die de scores op de 8 subtests het best samenvatten aan de hand van factoranalyse. Dit heeft als voordeel dat we minder regressieanalyses moeten uitvoeren en dat de resultaten stabieler zijn.

Factoranalyse op subtestscores van Vlamingen (θ) toont dat voor Vlamingen de 8 vaardigheidsscores kunnen samengevat worden in één factor die 52% van de variantie in de scores verklaart (zie Tabel 5.1). Aangezien alle subtests er op laden kunnen we de factor interpreteren als algemene intelligentie. Factor analyse op scores van allochtonen resulteert in twee factoren die samen 65% van de variantie in de scores verklaren. Zoals blijkt uit Tabel 5.1 correleert F2 vooral sterk met de twee verbale subtests (woordrelaties en woordanalogieën) terwijl F1 sterk correleert met de scores op de overblijvende 6 tests. In tegenstelling tot Vlamingen is bij allochtonen een onderscheid verbale en algemene intelligentie dus zinvol.

Tabel 5.2 toont het resultaat van stapsgewijze regressie per groep met als criterium de scores op de factoren en als predictoren de beschikbare achtergrondvariabelen.
Enkel de variabelen die in minstens 1 model significant bijdragen worden in de Tabel getoond. Als één van de niveaus van een polytome variabele significant is worden de andere niveaus ook opgenomen in het model.

Tabel 5.1 Factorladingen van subtests op factoren voor allochtonen en Vlamingen

	Vlamingen	allochtonen	
subtest	F1	F1	F2
woordrelaties	.68	.14	.89
woordanalogieën	.76	.26	.87
cijferreeksen	.80	.67	.40
komponenten	.74	.75	.20
kontroleren	.54	.77	.01
rekenvaardigheid	.69	.74	.22
spiegelbeelden	.73	.57	.46
exclusie	.79	.71	.28
eigenwaarde	4.2	4.0	1.1
%verklaarde variantie	52%	51%	14%

Tabel 5.2 Gestandaardiseerde regressiegewichten van achtergrondvariabelen per groep en per factor

	Allochtonen				Vlamingen	
Variabelen	F1		F2		F1	
	met TAAL SCORE	zonder TAAL SCORE	met TAAL SCORE	zonder TAAL SCORE	met TAAL SCORE	zonder TAAL SCORE
BASIS	-0.04	-0.08		-0.04	-0.17**	-0.22**
LAGSEC	-0.04	-0.06		-0.11*	-0.04	-0.05
HOGNTUNIV	0.10	0.09		0.03	0.20**	0.15*
UNIV	0.17**	0.18**		0.13*	0.18**	0.18**
TAALSCORE	0.21**	/	0.55**	/	0.34**	/
MCERV						0.13*
IMMIV7				0.07		
IMMIN7				-0.21*		
VERBLIJF			0.21**	0.20*		
MOENED			0.20**	0.26**		
Verkl variantie	**0.10**	**0.06**	**0.55**	**0.33**	**0.24**	**0.14**

* p<.05; **p<.01

In vergelijking met andere achtergrondvariabelen heeft de variabele taalscore een belangrijke bijdrage in de regressie. Dit is nog meer uitgesproken wanneer we voor de groep allochtonen de scores op de verbale factor (F2) trachten te verklaren. Dit sterke verband tussen F2 en de score op de taaltest is niet zo verwonderlijk omdat het in beide gevallen om een rechtstreekse meting van taalkennis gaat waar bovendien dezelfde testspecifieke factoren een rol kunnen spelen (motivatie, werken onder tijdsdruk, ..). Omwille van dit speciale statuut van de variabele TAALSCORE rapporteren we steeds een analyse met en zonder deze variabele.

Tabel 5.2 toont dat voor zowel Vlamingen als allochtonen de variabelen DIPLOMA en TAALSCORE een klein deel van variantie in F1 verklaren, respectievelijk 10% en 24%. Wanneer we TAALSCORE niet in de analyse opnemen, wordt nog maar 6% en 14 % van de variantie verklaard. DIPLOMA blijkt echter meer samen te hangen met F1 bij Vlamingen dan bij allochtonen. Het overgrote deel van de variantie binnen F1 kan echter niet verklaard worden op basis van de door ons geregistreerde persoonsvariabelen. Allochtonen die na hun 7 jaar zijn geïmmigreerd scoren lager dan allochtonen die hier geboren zijn.

De scores van allochtonen op de verbale factor F2 kunnen relatief beter verklaard worden dan de scores van allochtonen op de algemene niet-verbale factor F1. Het model waarin TAALSCORE wordt opgenomen verklaart een aanzienlijk deel van de variantie in F2, namelijk 55%. We stellen vast dat allochtonen hoger scoren op de verbale subtests naarmate ze hoger scoren op de taaltest, naarmate ze reeds langer in België verblijven en als ze Nederlands als moedertaal hebben. Als de variabele TAALSCORE niet wordt opgenomen in het model dan daalt het percentage verklaarde variantie tot 33% en dan zijn naast VERBLIJF en MOENED ook de variabelen DIPLOMA en IMMIGRATIE significant. Na controle voor de andere variabelen in het model blijkt dat men hoger scoort op F2 als men een hoger diploma heeft: allochtonen met een universitair diploma scoren hoger dan allochtonen met een diploma hoger secundair onderwijs (basiscategorie) en allochtonen met een diploma lager secundair onderwijs scoren lager dan allochtonen met een diploma hoger secundair onderwijs.

Naast het verklaren van de belangrijkste factoren in de 8 subtests kunnen we ook nagaan of specifieke bevindingen uit de literatuur van het intelligentie onderzoek bevestigd kunnen worden. Meerbepaald zijn er in de literatuur specifieke geslachts- en leeftijdsverschillen gevonden voor bepaalde soorten intelligentie. De internationale onderzoeksresultaten betreffende verschillen tussen mannen en vrouwen op cognitieve tests tonen dat vrouwen hogere scores behalen op het gebied van "verbal fluency" en "perceptual speed" en mannen hoger scoren op het gebied van "mathematics" (Hines, 1990; Born, Bleichrodt & Van der Flier, 1987). Volgens Kaufman & Horn (1996) en Born, Bleichrodt & Van der Flier (1987) is er ook een verband tussen leeftijd en gekristalliseerde en vloeiende intelligentie: De gekristalliseerde intelligentie neemt toe tot ongeveer 30 jaar en blijft vervolgens redelijk constant tot ongeveer 60 jaar, terwijl vloeiende intelligentie vanaf ongeveer 18 jaar langzaam afneemt.

Voor de subtests van de MCT-M impliceert dit het volgende:

- Mannen zouden beter moeten presteren op de subtest Rekenvaardigheid dan vrouwen
- Vrouwen zouden beter moeten scoren op de subtest Kontroleren
- Jongere kandidaten zouden beter moeten scoren op de subtests die vloeiende intelligentie meten (Exclusie, Komponenten, Spiegelbeelden en Kontroleren) dan oudere kandidaten

Deze hypothesen worden getoetst in een stapsgewijze regressie op de totale groep, waarbij de achtergrondvariabelen LEEFTIJD en VROUW in de analyse worden

opgenomen samen met nog enkele andere variabelen die voor beide groepen van toepassing zijn (DIPLOMA, TESTERV en MCERV). Verder wordt een nieuwe variabele opgesteld, namelijk ALLOCHT. Dit is de variabele die weergeeft of de kandidaat een allochtoon (ALLOCHT=1) is of niet (ALLOCHT=0).

Tabel 5.3 toont het resultaat van de stapsgewijze regressie per subtest met als criterium de subtestscores (θ) voor de verschillende subtests en als predictoren bovenvermelde achtergrondvariabelen. Enkel de variabelen die in minstens 1 model significant bijdragen worden in de Tabel getoond. Als één van de niveaus van een polytome variabele significant is, dan worden de andere niveaus ook opgenomen in het model.

Tabel 5.3 Ongestandaardiseerde regressiegewichten van achtergrondvariabelen voor de 8 MCT-M-subtests

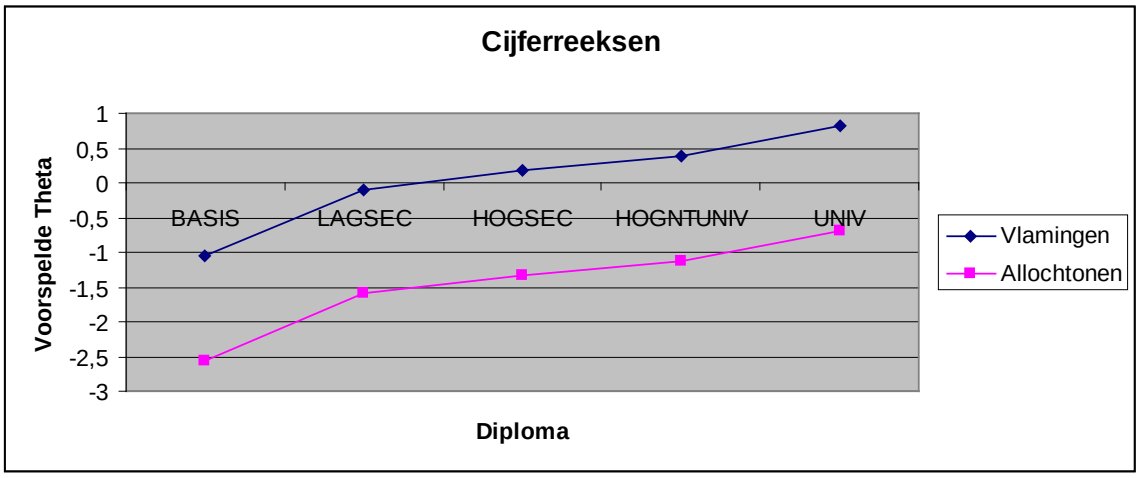
Variabele	Θ_{Cijfer}	$\Theta_{Exclusie}$	Θ_{Komp}	Θ_{Kontr}	Θ_{Rekenv}	$\Theta_{Spiegel}$	Θ_{Woana}	Θ_{Worel}
INTERCEPT	0.18	-0.003	0.04	-0.07	0.61**	0.10	0.11	-0.12
ALLOCHT	-1.51**	-1.03**	-0.87**	-0.78**	-1.84**	-1.85**	-1.92**	-2.49**
LEEFTIJD		-0.24**	-0.26**		0.24*	-0.35**		
VROUW				0.85**	-0.43*			
BASIS	-1.24**	-0.32*	-0.44	-1.45**	-1.21**	-0.97**	-1.06**	-1.14**
LAGSEC	-0.27	-0.02	-0.02	-0.30	-0.18	-0.05	-0.22	-0.25
HOGNTUNIV	0.20	0.37**	0.23	0.59	1.01**	0.28	0.44	0.73**
UNIV	0.65**	0.43**	0.61**	0.60	1.08**	0.44	1.00*	1.71**
BASIS*ALLOCHT							0.80*	0.64
LAGSEC*ALLOCHT							0.15	0.03
HOGNTUNIV*ALLOCHT							-0.42	-1.03**
UNIV * ALLOCHT							-0.83	-1.81**
R^2	0.27	0.27	0.11	0.07	0.17	0.17	0.41	0.64

Het geslacht van de kandidaat blijkt inderdaad enkel voor de subtests Kontroleren en Rekenvaardigheid een differentiërende rol te spelen: vrouwen zijn perceptueel sneller dan mannen en mannen zijn beter in het oplossen van rekenkundige problemen.
Wat het verschil in leeftijd betreft voor onze steekproef: Oudere mensen presteren minder goed op de subtests Exclusie, Komponenten en Spiegelbeelden. Deze drie subtests doen vooral beroep op vloeiende intelligentie, terwijl de andere tests (met uitzondering van Kontroleren) meer beroep doen op gekristalliseerde intelligentie. Deze resultaten bevestigen de bevindingen uit de literatuur.

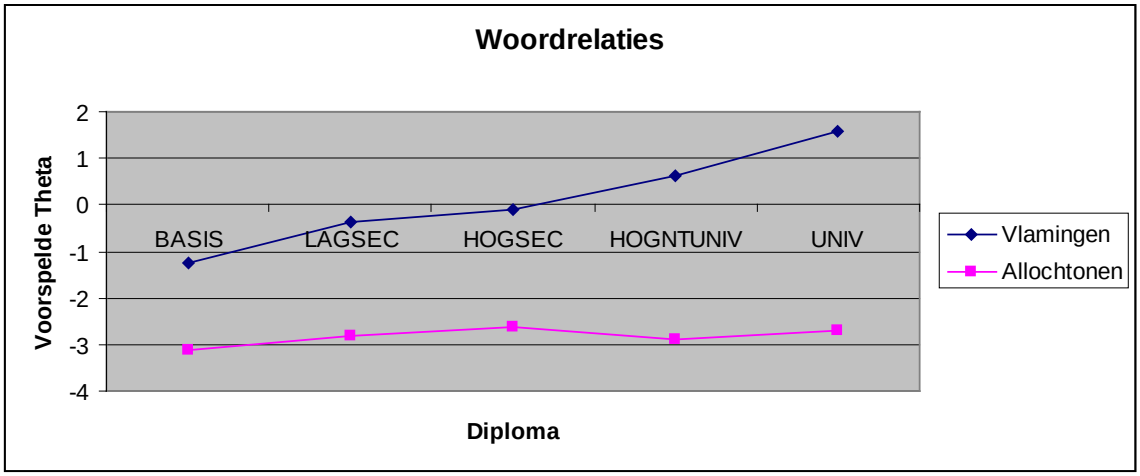
Verder toont Tabel 5.3 dat de achtergrondkenmerken bij sommige tests veel variantie in vaardigheidscores verklaren (Woordrelaties (64%), Woordanalogieën (41%)), terwijl ze bij andere tests relatief weinig bijdragen (Kontroleren (7%), Komponenten (11%)). Vooral de achtergrondvariabelen ALLOCHT en DIPLOMA zijn gerelateerd aan de vaardigheden die nodig zijn voor alle acht subtests.

Uit Tabel 5.3 blijkt dat er voor alle niet-verbale subtests een gelijkaardig hoofdeffect is van de variabele DIPLOMA. Personen met een hoger diploma behalen gemiddeld hogere

scores. Figuur 5.1 illustreert dit voor de subtest cijferreeksen. Voor de verbale subtests blijkt dat er niet alleen een hoofdeffect is van DIPLOMA, maar ook een significant interactie effect, wat wil zeggen dat het verband tussen DIPLOMA en scores op de verbale test anders is voor allochtonen en Vlamingen. De interactie effecten zijn tegengesteld aan het hoofdeffect (negatiever voor de hogere diploma's). Dit betekent concreet dat Vlamingen met een hoger diploma hogere testscores halen en dat voor allochtonen dit niet het geval is. Dit wordt geïllustreerd in Figuur 5.2 voor de subtest Woordrelaties. Figuur 5.2 laat zien dat de voorspelde scores voor allochtonen rond -2.8 schommelen, ongeacht het behaalde diploma. Merk wel op dat er binnen elke scholingsgraad zowel bij allochtonen als bij Vlamingen een tamelijk grote spreiding is van de vaardigheid en dat de Figuur slechts een voorspelling geeft van de gemiddelde scores.



Figuur 5.1 De voorspelde scores van Vlamingen en allochtonen op de subtest Cijferreeksen volgens diploma (gebaseerd op model in Tabel 5.3)



Figuur 5.2 De voorspelde scores van Vlamingen en allochtonen op de subtest Woordrelaties volgens diploma (gebaseerd op model in Tabel 5.3)

Hoofdstuk 6: Analyses ABL en SELOR

Zoals in Hoofdstuk 2 al werd vermeld, hebben we voor de tests van SELOR en ABL niet voldoende data om DIF betrouwbaar op te sporen. We beperken ons daarom bij deze tests tot het vergelijken van de totaalscores voor de verschillende groepen. Tabel 2.4 beschrijft de karakteristieken van de datasets van SELOR en ABL. De betrouwbaarheid van de tests is altijd hoog (α hoger dan .85). Uitzonderingen zijn NUMVA en de relatief korte test LOGDED.
Verder stellen we vast dat de betrouwbaarheid van de meeste tests ongeveer even hoog is voor de verschillende groepen.

Tabel 6.2 geeft een samenvatting van de gemiddelden en standaarddeviaties per groep voor SELOR- en ABL-tests. Bij de SELOR-tests LOGDED en ANAVERB scoren zowel Vlamingen als niet-Vlaamse EU kandidaten significant beter dan allochtonen. Bij CODES scoort enkel de Vlaamse groep significant beter dan de allochtone groep en is er geen significant verschil tussen niet-Vlaamse EU kandidaten en allochtonen. Bij NUMVA zijn er geen significante verschillen in gemiddelden tussen groepen.

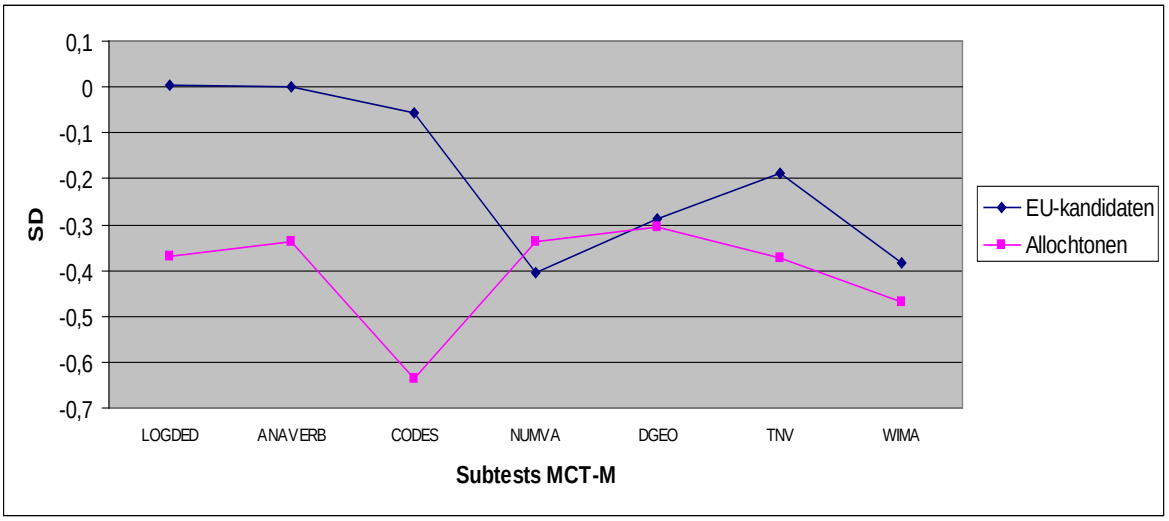
Voor de tests van ABL stellen we vast dat Vlamingen beter presteren op WIMA vergeleken met allochtonen en niet-Vlaamse EU kandidaten. Op TNV scoren Vlamingen enkel beter dan de allochtone steekproef. Een merkwaardig resultaat is dat bij DGEO Vlamingen hoger scoren dan niet-Vlaamse EU kandidaten, maar dat er geen significant verschil in gemiddelden blijkt te zijn tussen Vlamingen en allochtonen. Waarschijnlijk is het verschil tussen Vlamingen en allochtonen niet significant omdat de standaardfout van de allochtone groep groter is dan de standaardfout bij niet-Vlaamse EU kandidaten.

Tabel 6.2 Gemiddelden en standaarddeviaties per groep voor SELOR- en ABL-tests

Test		Vlamingen			niet-Vlaamse EU-kandidaten			Allochtonen		
	Max. score	Aantal Kandi-daten	Gemid. score	SD	Aantal Kandi-daten	Gemid. score	SD	Aantal Kandi-daten	Gemid. score	SD
LOGDED	22	2838	13,05 [a]	3,64	122	13,06[b]	3,69	79	11,70 [ab]	3,27
ANAVERB	100	2968	66,72 [a]	13,94	128	66,75 [b]	14,2	80	62,00 [ab]	12,60
CODES	74	877	41,72 [a]	14,86	20	40,86	12,1	16	32,25 [a]	18,57
NUMVA	38	1220	21,98	5,3	31	19,84	6,26	26	20,19	5,81
DGEO	40	862	21,78[a]	6,57	77	19,88 [a]	5,88	43	19,77	6,39
TNV	50	862	33,13 [a]	7,54	77	31,71	7,03	43	30,33 [a]	8,04
WIMA	23	862	12,16 [ab]	5,73	77	9,96[b]	5,16	43	9,47 [a]	5,51

a, b,= significant verschillend van elkaar

Om de relatieve score-verschillen tussen Vlamingen enerzijds en allochtonen/ niet-Vlaamse EU-kandidaten op de tests van SELOR en ABL te verduidelijken worden in figuur 6.1 de verschillen per test in standaarddeviaties van de Vlaamse groep weergegeven. Uit figuur 6.1 blijkt dat de verschillen tussen Vlamingen en allochtonen voor bijna alle tests (met uitzondering van de NUMVA-test) groter zijn dan de verschillen tussen Vlamingen en niet-Vlaamse EU kandidaten. We kunnen dit zien in de Figuur omdat de verschillen tussen Vlamingen en niet-Vlaamse EU kandidaten uitgedrukt in SDs van de Vlaamse groep dichter bij nul liggen dan de verschillen tussen Vlamingen en allochtonen. Bij de SELOR-tests LOGDED, ANAVERB en CODES is er bijna geen verschil in gemiddelden tussen Vlamingen en niet-Vlaamse EU kandidaten, terwijl er bij de ABL-tests DGEO en WIMA wel duidelijke (en zelfs significante) verschillen zijn in gemiddelden tussen Vlamingen en niet-Vlaamse EU kandidaten. Het verschil in gemiddelden tussen Vlamingen en allochtonen varieert tussen .31 en .64 standaarddeviaties en is voor bijna alle tests (met uitzondering van NUMVA en DGEO) significant. Het verschil tussen Vlamingen en niet-Vlaamse EU kandidaten varieert tussen 0 en .40 en is enkel bij DGEO en WIMA significant.



Figuur 6.1 Verschillen in gemiddelde testscores tussen enerzijds Vlamingen en anderzijds allochtonen/ niet-Vlaamse EU kandidaten uitgedrukt in standaarddeviaties van de Vlaamse groep.

Hoofdstuk 7: Samenvatting van onderzoeksresultaten en beleidsaanbevelingen

7.1 Samenvatting van onderzoeksresultaten MCT-M

In dit rapport werd voor acht verschillende types van intelligentietests (MCT-M) onderzocht of ze discriminerend zijn voor allochtonen versus Vlamingen. Discriminatie werd onderzocht op het niveau van individuele items en op het niveau van de test als geheel aan de hand van het concept Differential Item Functioning (DIF) uit de itemresponstheorie. Daarnaast werd voor de subtests met significante DIF onderzocht in welke mate DIF in individuele items kon verklaard worden op basis van itemkenmerken. In wat volgt geven we een samenvatting van de wetenschappelijke resultaten van het onderzoek. Daarna formuleren we hierbij aansluitende beleidsaanbevelingen.

7.1.1 Gemiddelde prestaties van Vlamingen en allochtonen

7.1.1.1 Verschillen in gemiddelde testscores

Tabel 7.1 geeft voor de onderzochte tests een overzicht van de voornaamste testkarakteristieken (grootte van de dataset, interne consistentie) en van de gemiddelde prestaties van Vlamingen en allochtonen. De interne consistentie van de tests (bepaald aan de hand van Cronbach's α) is hoog (α>.80) en verschilt zeer weinig voor Vlamingen en allochtonen. Dit wil zeggen dat alle tests tamelijk betrouwbaar zijn, zowel voor Vlamingen als voor allochtonen.

De gemiddelde prestaties van Vlamingen en allochtonen worden weergegeven door het gemiddelde van de vaardigheid θ in elke groep (μ_θ). Voor Vlamingen wordt de gemiddelde vaardigheid op voorhand gelijk gesteld aan nul om het nulpunt van de latente schaal vast te leggen. Voor allochtonen wordt de gemiddelde vaardigheid bepaald in de veronderstelling dat de items in de twee groepen gemiddeld even moeilijk zijn (assumptie van gelijke populatiegemiddelden). De spreiding van de vaardigheid in elke groep (SD(θ)) wordt bepaald op basis van de gegevens.

Tabel 7.1 toont dat allochtonen gemiddeld lagere scores behalen dan Vlamingen voor alle subtests. Figuur 7.1 toont de omvang van de gevonden verschillen uitgedrukt in standaarddeviaties van de vaardigheidsverdeling van Vlamingen. We zien dat het verschil tussen Vlamingen en allochtonen het grootst is op de verbale subtests (voor woordrelaties meer dan 2 SD en voor woordanalogieën 1.5 SD). Bij de overige subtests varieert het verschil tussen .37 en .87 SD. Het gemiddelde verschil bedraagt .89 SD, wat in de lijn ligt van bevindingen die eerder gerapporteerd werden in de literatuur (Van den Berg, 2001).

Om de verschillen tussen Vlamingen en allochtonen te vertalen in termen van succeskansen berekenen we de succeskans van een gemiddelde allochtoon voor een item met gemiddelde moeilijkheid voor Vlamingen (dit wil zeggen dat een gemiddelde

Vlaming (μ_0=0) 50% kans heeft om dit item juist te beantwoorden). Zoals blijkt uit de laatste kolom van Tabel 7.1 variëren de succeskansen van de gemiddelde allochtoon van .07 tot .32 wat aanzienlijk lager is dan .50. Voor de verbale subtests woordrelaties en woordanalogieën zijn de succeskansen van de gemiddelde allochtoon zelfs meer dan 4 keer kleiner.

Men kan op basis van Tabel 7.1 concluderen dat allochtonen vooral slechter presteren dan Vlamingen op de verbale subtests. Deze conclusie is gerechtvaardigd omdat alle subtests werden aangeboden aan dezelfde steekproef van Vlamingen en allochtonen. De precieze omvang van het geobserveerde verschil in gemiddeld presteren is echter niet noodzakelijk generaliseerbaar naar de volledige populatie van allochtonen en Vlamingen en naar andere onderzoeken. De steekproeven in dit onderzoek zijn immers tamelijk klein en niet representatief. Bovendien blijkt uit de literatuur dat de scores van allochtonen op verbale tests sterk bepaald worden door de verblijfsduur, het tijdstip waarop men geïmmigreerd is, en of men eerste of tweede generatie allochtoon is. Een herhaling van het onderzoek met een andere samenstelling van de steekproef van allochtonen kan dus sterk verschillende resultaten geven. We kunnen hierbij opmerken dat het sterke verschil in gemiddeld presteren op de verbale subtests zeer groot is omdat de steekproef vooral eerste generatie allochtonen bevat.

7.1.1.2 Verklaren van verschillen in gemiddelde testscores op basis van achtergrondvariabelen

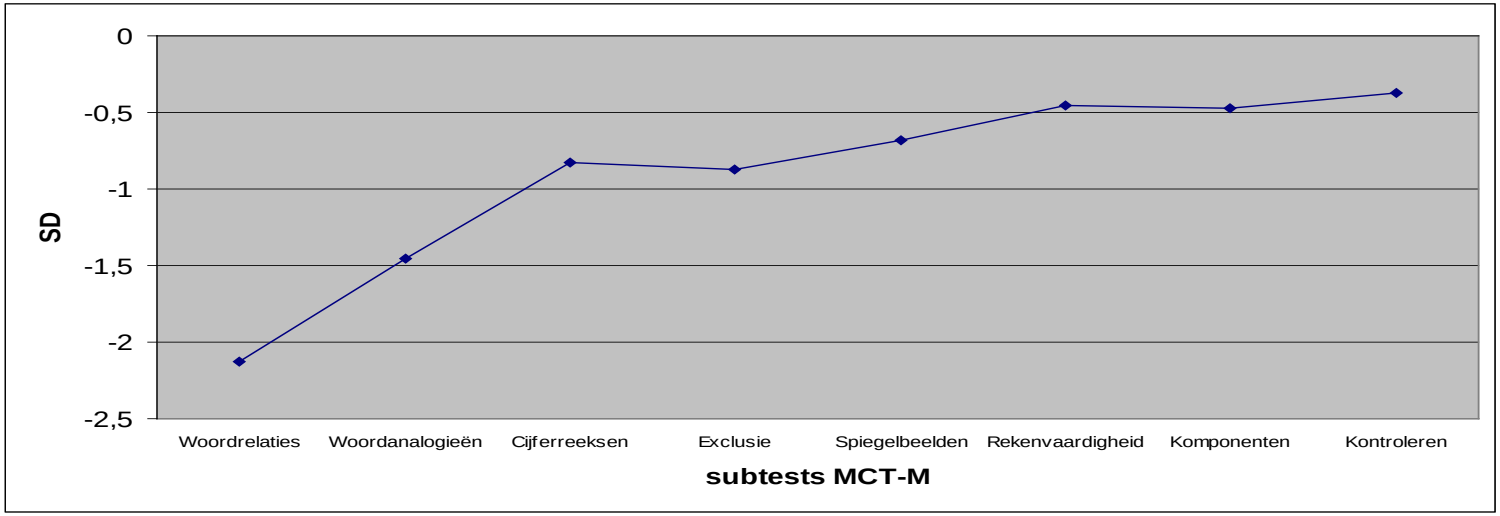
Om te onderzoeken in welke mate de scores van allochtonen en Vlamingen op de subtests van de MCT-M kunnen verklaard worden op basis van achtergrondvariabelen gebruiken we multiple regressie met achtergrondvariabelen als predictoren. Als afhankelijke variabele gebruiken we een minimaal aantal factoren die de scores op de 8 subtests het best samenvatten op basis van factoranalyse. Factoranalyse toont dat de scores van Vlamingen op de 8 subtests kunnen gevat worden met 1 factor (F1) die kan geïnterpreteerd worden als algemene intelligentie. De scores van allochtonen op de 8 subtests kunnen gereduceerd worden tot 2 factoren (F1 en F2). De eerste factor (F1) correleert sterk met alle subtests behalve met de verbale tests (woordrelaties en woordanalogieën) en de tweede factor (F2) correleert alleen met de verbale tests. De bevinding dat bij Vlamingen alle tests (en dus ook de verbale tests) op dezelfde factor laden wil zeggen dat ze allemaal een hoog "algemene-intelligentie" gehalte hebben. Dat de verbale tests bij allochtonen een aparte factor vormen wil zeggen dat de scores op deze tests waarschijnlijk ook bepaald worden door schoolse kennis en vaardigheden (van den Berg, 2001).

Regressie van factorscores op achtergrondvariabelen toont dat de scores op algemene-intelligentie factoren (F1 bij Vlamingen en allochtonen) maar zeer beperkt verklaard kunnen worden. Enkel het behaalde diploma en de score op de taaltest spelen een rol in de zin dat een hoger diploma of een hogere score op de taaltest gemiddeld leidt tot hogere scores op F1.

Tabel 7.1 Steekproefomvang (N), interne consistentie (α), gemiddelde vaardigheid (μ_θ) van MCT-M tests voor allochtonen en Vlamingen. Aantal items per test en succeskans van een gemiddelde allochtoon op een unbiased item met gemiddelde moeilijkheid voor Vlamingen (β_i=0).

Test	aantal items	Vlamingen				allochtonen				succeskans gemiddelde allochtoon op unbiased item met β_i=0
		N	α	μ_θ	SD(θ)	N	α	μ_θ	SD(θ)	
REKENVAARDIGHEID	30	241	.92	0	2.6	289	.93	-1.19**	2.5	0.23
KOMPONENTEN	30	240	.89	0	1.6	285	.91	-.73**	1.6	0.32
WOORDRELATIES	45	241	.89	0	1.2	288	.88	-2.57**	1.2	0.07
CIJFERREEKSEN	30	242	.86	0	1.5	286	.85	-1.23**	1.3	0.23
KONTROLEREN	36	239	.97	0	2.9	288	.97	-1.1**	3.0	0.25
SPIEGELBEELDEN	30	236	.96	0	2.5	285	.96	-1.66**	2.4	0.16
WOORDANALOGIEËN	30	243	.87	0	1.3	283	.90	-1.88**	1.3	0.13
EXCLUSIE	30	242	.81	0	1.1	286	.84	-.96**	1.1	0.28

Noot: * p<.05; **p<.01; De gegevens voor de subtest Kontroleren zijn gebaseerd op items 40-75



Figuur 7.1 Verschillen in gemiddelde vaardigheid (μ_θ) tussen Vlamingen en allochtonen uitgedrukt in standaarddeviaties van de vaardigheidsverdeling voor Vlamingen
Noot: De gegevens voor de subtest Kontroleren zijn gebaseerd op items 40-75.

De scores van allochtonen op F2 kunnen relatief beter verklaard worden op basis van achtergrondvariabelen. Naast de score op de taaltest en het behaalde diploma speelt ook nog het aantal jaren in België, de moedertaal en de leeftijd bij immigratie een rol. Men haalt hogere scores op de verbale factor F2 als men Nederlands als moedertaal heeft, en naarmate men langer in België is, en men scoort lager als men pas na het zevende levensjaar naar België is gekomen.

Naast het verklaren van de belangrijkste factoren in de 8 subtests kunnen we ook nagaan of specifieke bevindingen uit de literatuur van het intelligentie onderzoek bevestigd kunnen worden. Meerbepaald zijn er in de literatuur specifieke geslachts- en leeftijdsverschillen gevonden voor bepaalde soorten intelligentie. De internationale onderzoeksresultaten betreffende verschillen tussen mannen en vrouwen op cognitieve tests tonen dat vrouwen hogere scores behalen op het gebied van "verbal fluency" en "perceptual speed" en mannen hoger scoren op het gebied van "mathematics" (Hines, 1990; Born, Bleichrodt & Van der Flier, 1987). Volgens Kaufman & Horn (1996) en Born, Bleichrodt & Van der Flier (1987) is er ook een verband tussen leeftijd en gekristalliseerde en vloeiende intelligentie: De gekristalliseerde intelligentie neemt toe tot ongeveer 30 jaar en blijft vervolgens redelijk constant tot ongeveer 60 jaar, terwijl vloeiende intelligentie vanaf ongeveer 18 jaar langzaam afneemt.

De resultaten van ons onderzoek bevestigen in grote lijnen de bevindingen uit de literatuur. Het geslacht van de kandidaat blijkt inderdaad enkel voor de subtests Kontroleren en Rekenvaardigheid een differentiërende rol te spelen: vrouwen zijn perceptueel sneller en mannen zijn beter in het oplossen van rekenkundige problemen. Wat het verschil in leeftijd betreft voor onze steekproef: Oudere mensen presteren minder goed op de subtests Exclusie, Komponenten en Spiegelbeelden. Deze drie subtests doen vooral beroep op vloeiende intelligentie, terwijl de andere tests (met uitzondering van Kontroleren) meer beroep doen op gekristalliseerde intelligentie.

7.1.2 Snelheid versus vaardigheid

De MCT-M bevat zowel zogenaamde "powertests" als "snelheidstests". Bij zuivere powertests primeert kennis en vaardigheid en zal men soms het juiste antwoord niet vinden/weten ook al heeft men zoveel tijd om na te denken als men wil. Bij zuivere snelheidstests zijn alle items ongeveer even moeilijk en kan men in principe altijd de juiste oplossing vinden als men tijd genoeg heeft. De subtests van de MCT-M hebben allemaal een tijdslimiet maar toch nemen ze elk een verschillende positie in op het power-snelheids continuüm

Om te onderzoeken in welke mate snelheid en vaardigheid van belang zijn om verschillen tussen Vlamingen en allochtonen in verschillende subtests te verklaren, hebben we het verband bestudeerd tussen drie variabelen die berekend kunnen worden voor elk item. Een eerste variabele, aangeduid als de globale proportie juist, is gelijk aan de proportie personen die het item juist ingevuld heeft. Als het item niet ingevuld is, wordt het fout gerekend. Een tweede variabele, aangeduid als de proportie ingevuld, is de proportie personen die het item ingevuld heeft. Een derde variabele, aangeduid als de proportie

juist gegeven ingevuld, is gelijk aan de proportie personen die het item juist heeft gegeven dat ze het hebben ingevuld.

Regressie van het verschil in globale proportie juist bij Vlamingen en allochtonen op het verschil in proportie ingevuld en op het verschil in de proportie juist gegeven ingevuld toont in de verschillende subtests het belang van snelheid en vaardigheid. Tabel 7.2 toont dat het verschil in globale proportie juist tussen Vlamingen en allochtonen verklaard wordt door zowel een verschil in vaardigheid, als een verschil in snelheid maar er is er een verschil in bijdrage van beide factoren, afhankelijk van de subtest. Sommige subtests meten vooral een verschil in snelheid (Rekenvaardigheid, Kontroleren, Komponenten). Voor deze tests zijn alle items ongeveer even moeilijk en is het dus vooral een kwestie van snel antwoorden. Andere tests meten voornamelijk een verschil in vaardigheid (Spiegelbeelden, Woordrelaties en Woordanalogieën). Bij deze tests met relatief moeilijke items is het vooral belangrijk dat men over de vaardigheid/kennis beschikt om deze items op te lossen. Tot slot zijn er tests waar zowel verschillen in snelheid als verschillen in vaardigheid een rol spelen (bijvoorbeeld, Exclusie, Cijferreeksen).

Uit Tabel 7.1 blijkt dat Vlamingen gemiddeld beter presteren op alle subtests. De verschillen zijn echter kleiner naarmate de test meer een pure snelheidstest is (Rekenvaardigheid, Komponenten, Kontroleren). Dit blijkt bijvoorbeeld uit de resultaten op de administratieve test Kontroleren waar het relatief kleine (maar significante) verschil in gemiddelde vaardigheid een puur snelheidseffect is. Anderzijds blijkt dat als kennis belangrijker wordt, het verschil in gemiddelde vaardigheid nog groter wordt. Bij de verbale tests waar voldoende kennis van Nederlands belangrijk is, zijn de verschillen in gemiddelde vaardigheid het grootst.

7.1.3 DIF in de moeilijkheidsgraden

Zoals blijkt uit Tabel 7.3 zijn er bij een aantal tests relatief veel items die DIF vertonen (Woordrelaties, Woordanalogieën en Exclusie), terwijl er bij andere tests weinig tot geen DIF optreedt (Cijferreeksen, Spiegelbeelden, Komponenten, Rekenvaardigheid, Kontroleren). De proportie van alle items die significante DIF vertonen op het 5% niveau varieert over tests van .00 tot .55. De verbale tests vertonen het meeste DIF, namelijk in de helft van alle items. Bij Cijferreeksen en Spiegelbeelden is er geen DIF in de items aanwezig. Als er DIF in de test aanwezig is, zoals bij de verbale subtests, dan is dit soms in het voordeel van allochtonen en soms in het voordeel van Vlamingen. Dit is een rechtstreeks gevolg van de methode die gebruikt wordt om het verschil in gemiddeld presteren (hoofdeffect) te bepalen (methode van gelijke populatiegemiddelden).

Omdat de statistische significantie van de DIF sterk bepaald wordt door de omvang van de geanalyseerde steekproeven evalueren we ook de praktische significantie van de DIF aan de hand van de verdeling van de absolute verschillen tussen IRFs van mannen en vrouwen. Uit de resultaten in Tabel 7.3 blijkt dat de praktische significantie van de DIF voor sommige tests tamelijk groot is.

Tabel 7.2 Bijdrage van verschil in vaardigheid en verschil in snelheid aan verschil in globale proportie juist bij Vlamingen en allochtonen.

Test	Verschil in vaardigheid		Verschil in snelheid		R^2
	gewicht	p-waarde	gewicht	p-waarde	
REKENVAARDIGHEID	0,10	0,0044	0,99	<,0001	0,97
KONTROLEREN	0,14	<,0001	0,96	<,0001	0,96
KOMPONENTEN	0,30	0,0003	0,79	<,0001	0,87
EXCLUSIE	0,61	<,0001	0,46	<,0001	0,91
CIJFERREEKSEN	0,56	0,0003	0,38	0,0078	0,76
WOORDANALOGIEÊN	0,79	<,0001	0,31	<,0001	0,98
SPIEGELBEELDEN	0,84	<,0001	-0,39	0,0001	0,79
WOORDRELATIES	0,95	<,0001	0,02	0,7214	0,91

Tabel 7.3 Proportie items per test die DIF vertonen volgens verschillende criteria

Test	Aantal Vlamingen	Aantal allochtonen	Proportie items met DIF op 5% niveau	Proportie van alle items met bepaalde waarde voor Mediaan van de absolute verschillen tussen IRFs van Vlamingen en allochtonen			Proportie van alle items met bepaalde waarde voor 97.5 percentiel van de absolute verschillen tussen IRFs van Vlamingen en allochtonen			
				Me<.05	.05<Me<.10	Me>.10	$P_{97.5}$<.10	.10<$p_{97.5}$<.15	.15<$p_{97.5}$<.20	$p_{97.5}$>.20
REKENVAARDIGHEID	241	289	0.07	0.83	0.17	0	0.47	0.33	0.13	0.07
KOMPONENTEN	240	285	0.07	0.93	0.07	0	0.86	0.07	0.07	0
WOORDRELATIES	241	288	**0.55**	0.40	0.22	0.38	0.27	0.09	0.13	0.51
CIJFERREEKSEN	242	286	0	1.00	0	0	0.80	0.17	0	0.03
KONTROLEREN	239	288	0.06	0.97	0.03	0	0.81	0.19	0	0
SPIEGELBEELDEN	236	285	0	0.93	0.07	0	0.93	0.07	0	0
WOORDANALOGIEËN	243	283	**0.50**	0.67	0.20	0.13	0.31	0.23	0.13	0.33
EXCLUSIE	242	286	**0.37**	0.67	0.33	0	0.40	0.33	0.20	0.07

Zo stellen we vast dat in de subtest Woordrelaties 38% van de items grote absolute verschillen vertonen (mediaan >.10) en dat maar 40% van de items een mediaan kleiner dan .05 heeft. Voor de overige tests geldt dat de mediaan van de absolute verschillen tussen IRFs van Vlamingen en allochtonen voor het grootste deel van de items wel kleiner is dan .05.

Verder blijkt dat in bepaalde tests voor een aantal items en voor een bepaald deel van de latente schaal de DIF sterk kan oplopen. Bijvoorbeeld voor Woordrelaties en Woordanalogieën is voor respectievelijk 51% en 33% van de items het 97.5 percentiel van de absolute verschillen tussen IRFs van Vlamingen en allochtonen groter dan .20. Met andere woorden, in ongeveer 51% van de Woordrelaties-items en in 33% van de Woordanalogieën-items zijn de verschillen tussen de succeskansen van Vlamingen en allochtonen op een klein stuk van de latente schaal (2.5%) minstens .20. Bij de subtest Exclusie zijn er 18 items waar de verschillen in succeskans van Vlamingen en allochtonen op 2.5% van de latente schaal groter dan .10 zijn. Op basis van de resultaten in de tabel kunnen we besluiten dat DIF vooral in de verbale tests (Woordrelaties en Woordanalogieën) voorkomt. Het is dus van belang om inzicht te krijgen in de factoren die verbale items moeilijker of gemakkelijker maken voor allochtonen zodat de constructvaliditeit van de verbale tests kan verbeterd worden.

7.1.4 Gezamenlijk effect van DIF in individuele items op de somscores

Voor alle onderzochte tests blijkt dat de DIF-effecten in individuele items elkaar opheffen en als dusdanig geen effect hebben op de (verwachte) somscores. Na correctie voor het verschil in gemiddelde vaardigheid is het verband tussen de latente variabele en de testscore voor alle onderzochte tests hetzelfde in beide groepen. Hierbij moeten echter twee kanttekeningen geplaatst worden.

Ten eerste is het van belang om op te merken dat dit resultaat sterk bepaald wordt door de manier waarop het verschil in gemiddelde vaardigheid bepaald is. De methode van gelijke populatiegemiddelden stelt dat de gemiddelde moeilijkheid van alle items dezelfde is in de twee groepen. Als de moeilijkheidsgraden in de twee groepen verschillen dan leidt deze methode in de regel tot zowel positieve als negatieve DIF. Een andere methode om het hoofdeffect te bepalen zou bijvoorbeeld kunnen leiden tot een minder groot verschil in gemiddeld presteren in het voordeel van Vlamingen en meer DIF-items in het voordeel van Vlamingen. Deze afhankelijkheid tussen het verschil in gemiddeld presteren en DIF is inherent aan het concept van Differential Item Functioning.

Ten tweede moet men er zich van bewust zijn dat als er veel tegengestelde DIF optreedt éénzelfde testscore iets anders kan betekenen voor Vlamingen en allochtonen. Ze kan immers het resultaat zijn van succes op verschillende verzamelingen van items die elk een tegengestelde DIF vertonen. Tests met veel DIF (verbale tests, Exclusie) meten dus gedeeltelijk verschillende aspecten in de twee groepen.

We kunnen besluiten dat voor tests met weinig DIF de somscores in beide groepen kunnen vergeleken worden, maar dat voor tests met veel DIF de scores in principe een verschillende betekenis kunnen hebben.

In de literatuur vindt men ook dat vooral de verbale tests meer DIF-items bevatten dan andere tests (Poortinga & Van der Flier, 1988). De effecten van DIF op de totaalscores zijn in de meeste studies wel beduidend groter en meestal in het nadeel van de allochtone groep (Ellis, 1990; Schmitt & Dorans, 1990; Te Nijenhuis, 1997). Enkel het onderzoek van van den Berg (2001) toont, net zoals onze studie, bijna geen effect van DIF op de totaalscore. Mogelijks kan dit verschil in resultaten verklaard worden doordat bij de ontwikkeling van de MCT-M reeds DIF onderzoek heeft plaatsgevonden en DIF items al verwijderd werden. De methode die gebruikt is om DIF op te sporen zou echter ook een rol kunnen spelen.

Voor de meeste tests blijkt dat het 95% betrouwbaarheidsinterval dat wordt afgebakend rond de somscores erg breed is. Bijvoorbeeld in de steekproef van Vlamingen bij Spiegelbeelden is de standaardmeetfout uit de klassieke testtheorie gelijk aan 1.79, wat wil zeggen dat een somscore 17 (berekend op 30 items) met 95% zekerheid ligt tussen 17-(1.96*1.79)=13.5 en 17+(1.96*1.79)=20.5. De standaardmeetfouten van de acht subtests variëren tussen 1.57 en 2.38. Dit zijn grote standaardmeetfouten, waardoor de betrouwbaarheidsintervallen rond de somscores veel overlap vertonen. Bij het ordenen van de kandidaten op basis van de somscore is het van belang hiermee rekening te houden. Kandidaten kunnen immers strikt genomen alleen geordend worden als de 95% betrouwbaarheidsintervallen rond hun somscores niet overlappen.

Een voordeel van het schatten van de vaardigheid met itemresponsmodellen in plaats van met klassieke testtheorie is dat de standaardfout van de persoonsparameter θ kleiner is voor delen van de latente schaal waar men veel informatie heeft over de vaardigheid (namelijk omdat er zich veel items bevinden) en groter is voor delen van de schaal waar men weinig informatie heeft over de te bepalen vaardigheid (namelijk omdat er zich weinig items bevinden). De standaardfout van de vaardigheid kan dus in principe nauwkeuriger geschat worden met methoden uit de itemresponstheorie zodat ook het strikt ordenen van personen nauwkeuriger kan gebeuren.

7.1.5 Verklaren van DIF

Bij het verklaren van DIF hebben we ons beperkt tot de subtests waar er veel DIF aanwezig was en waarvan de items in aanmerking kwamen voor een inhoudelijke analyse. Bijgevolg werden enkel de verbale subtests Woordrelaties en Woordanalogieën grondig bestudeerd.

Uit het verklarend onderzoek op de subtests Woordrelaties en Woordanalogieën kunnen volgende conclusies getrokken worden:

- Items zijn voor beide groepen moeilijker wanneer de woorden van het item minder voorkomen in gesproken of geschreven Nederlands en/of naarmate de woorden abstracter zijn.
- Items zijn voor allochtonen moeilijker wanneer ze woorden bevatten die minder vaak voorkomen in het dagelijks taalgebruik en/of als de woorden typisch zijn voor het Nederlands (en dus niet kunnen afgeleid worden van een andere taal). Bij de test Woordrelaties blijkt het zoeken naar synoniemen moeilijker voor allochtonen dan het zoeken naar tegenstellingen.

7.2 Samenvatting van onderzoeksresultaten ABL en SELOR

Voor de tests van SELOR en ABL beperkt het onderzoek zich tot het vergelijken van de totaalscores voor verschillende groepen. Er worden drie groepen vergeleken: Vlamingen, allochtonen en niet-Vlaamse EU kandidaten. Tabel 7.4 geeft een overzicht van de voornaamste testkarakteristieken (grootte van de dataset, interne consistentie) en van de gemiddelde prestaties van Vlamingen, niet-Vlaamse-EU kandidaten en allochtonen. De interne consistentie van de tests (bepaald aan de hand van coëfficiënt α) is voor de meeste tests hoog (>. 80), met uitzondering van de relatief korte test LOGDED. De interne consistentie verschilt zeer weinig voor de verschillende groepen, wat wil zeggen dat deze tests tamelijk betrouwbaar zijn voor zowel Vlamingen, niet-Vlaamse EU-kandidaten als allochtonen.

Bij de SELOR-tests presteren Vlamingen gemiddeld altijd beter dan allochtonen, met uitzondering van NUMVA, waar er geen significante verschillen tussen de groepen worden vastgesteld. Dat het verschil bij NUMVA niet significant is, heeft natuurlijk ook te maken met het relatief kleine aantal allochtone en niet-Vlaamse-EU kandidaten waardoor de test minder snel significant is. Voor LOGDED en ANAVERB hebben niet-Vlaamse-EU kandidaten gemiddeld ook hogere scores dan allochtonen.

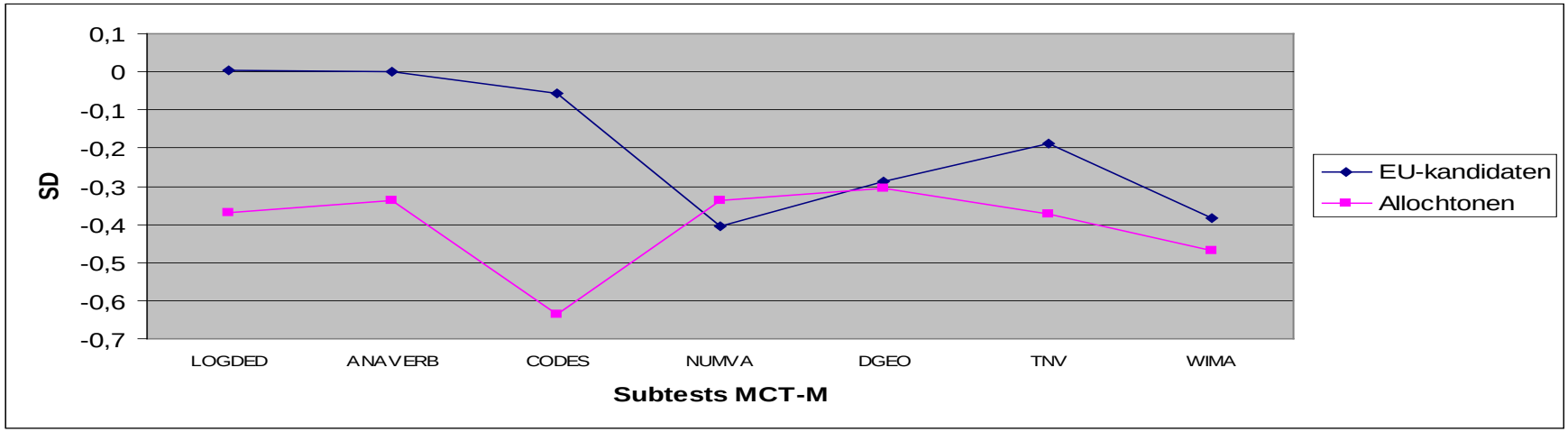
Op de tests TNV en WIMA van ABL presteren Vlamingen gemiddeld significant beter dan allochtonen en op de tests DGEO en WIMA scoren zij gemiddeld beter dan niet-Vlaamse-EU kandidaten. Tussen allochtonen en niet-Vlaamse-EU kandidaten is er geen significant verschil in gemiddelde testscores.

Om de relatieve verschillen tussen enerzijds Vlamingen en anderzijds allochtonen/ niet-Vlaamse-EU kandidaten op de tests van SELOR en ABL te visualiseren worden in figuur 7.2 de verschillen per test weergegeven uitgedrukt in standaarddeviaties van de Vlaamse groep. Het verschil in gemiddelden tussen Vlamingen en allochtonen varieert tussen .31 en .64 standaarddeviaties en is voor bijna alle tests (met uitzondering van de tests NUMVA en DGEO) significant. Het verschil tussen Vlamingen en niet-Vlaamse-EU kandidaten varieert tussen 0 en .40 en is enkel voor de tests DGEO en WIMA significant.

Tabel 7.4 Kenmerken van testgegevens bij SELOR en ABL: aantal kandidaten (N), interne consistentie (α), gemiddelde testscore (μ) en standaarddeviatie van testscores (SD)

Organisatie	Test		Vlamingen				Niet-Vlaamse-EU kandidaten				Allochtonen			
		aantal items	N	α	μ	SD	N	α	μ	SD	N	α	μ	SD
SELOR	LOGDED	22	2838	0,76	13,05 [a]	3,64	122	0,76	13,06[b]	3,69	79	0,70	11,70 [ab]	3,27
SELOR	ANAVERB	100	2968	0,93	66,72 [a]	13,94	128	0,93	66,75 [b]	14,2	80	0,91	62,00 [ab]	12,6
SELOR	CODES	74	877	0,94	41,72 [a]	14,86	20	0,91	40,86	12,1	16	0,97	32,25 [a]	18,57
SELOR	NUMVA	38	1220	0,80	21,98	5,3	31	0,85	19,84	6,26	26	0,82	20,19	5,81
ABL	DGEO	40	862	0,88	21,78[a]	6,57	77	0,85	19,88 [a]	5,88	43	0,87	19,77	6,39
ABL	TNV	50	862	0,88	33,13 [a]	7,54	77	0,86	31,71	7,03	43	0,89	30,33 [a]	8,04
ABL	WIMA	23	862	0,88	12,16 [ab]	5,73	77	0,84	9,96[b]	5,16	43	0,87	9,47 [a]	5,51

a, b, … = significant verschillend van elkaar



Figuur 7.2 Verschillen in testscores tussen enerzijds Vlamingen en anderzijds allochtonen/niet-Vlaamse-EU kandidaten uitgedrukt in standaarddeviaties van de Vlaamse groep.

Merk op dat verschillen in testscores (uitgedrukt in standaarddeviaties van testscores voor Vlamingen) bij de onderzochte tests van SELOR en ABL in het algemeen kleiner zijn dan bij de MCT-M test van VDAB. Dit kan waarschijnlijk verklaard worden door een verschillende samenstelling van de steekproef van allochtonen. In het onderzoek van VDAB bestond de groep allochtonen vooral uit eerste generatie allochtonen terwijl bij SELOR/ABL misschien meer tweede generatie allochtonen deelnamen. Analyse van de achtergrondkenmerken van de allochtone groep bij SELOR en ABL (nog uit te voeren) zal hierover uitsluitsel geven.

7.3 Beleidsaanbevelingen

Op basis van de resultaten van het onderzoek kunnen verschillende aanbevelingen gemaakt worden. Deze kunnen gericht zijn aan verschillende instanties, zoals de producenten van tests, de selecteurs die de tests gebruiken en de overheid die een gelijkekansenbeleid tracht te voeren. De aanbevelingen worden niet op de ene of de andere instantie toegespitst. Omdat de overheid ook een rol speelt in de adviezen aan de testproducenten en selecteurs, en omdat omgekeerd de testproducenten en selecteurs betrokken dienen te worden in de uitvoering van de beleidsaanbevelingen aan de overheid, worden de aanbevelingen in hun algemeenheid geformuleerd.

1. Het onderzoek wijst uit dat allochtonen op alle onderzochte tests gemiddeld slechter scoren dan Vlamingen (behalve bij NUMVA waar de steekproef allochtonen waarschijnlijk te klein is om significante resultaten te bekomen). De verschillen in gemiddelde vaardigheid uitgedrukt in standaarddeviaties van de testscores voor Vlamingen zijn voor de meeste subtests tussen .30 SD en .80 SD lager voor allochtonen. Voor de verbale subtests van de MCT-M zijn de verschillen evenwel nog groter (1.4 en 2.1 voor respectievelijk woordanalogieën en woordrelaties). Merk op dat in dit onderzoek vooral eerste generatie allochtonen deelnamen.

Factoranalyse van subtestscores per groep toont dat de 8 subtests van de MCT-M een verschillende factorstructuur hebben voor Vlamingen en allochtonen. In de Vlaamse steekproef kunnen alle subtestscores verklaard worden op basis van 1 algemene-intelligentie factor, terwijl bij allochtonen twee factoren nodig zijn: een algemene intelligentie-factor en een verbale factor. Dit betekent dat voor allochtonen de verbale subtests van de MCT-M niet alleen algemene intelligentie meten maar ook schoolse kennis en vaardigheden (meerbepaald kennis van Nederlands). Verder laten de resultaten van ons onderzoek zien dat scores op de algemene-intelligentie factoren in beperkte mate samenhangen met opleidingsniveau. De scores van allochtonen op de verbale factor daarentegen kunnen iets beter verklaard worden en hangen samen met opleidingsniveau, verblijfsduur, leeftijd bij immigratie, de moedertaal, de score op een taaltest voor Nederlands. Kennis van het Nederlands zoals gemeten met de taaltest blijkt de belangrijkste factor om te verklaren waarom bepaalde allochtonen slechter scoren.

Als allochtonen gemiddeld vooral slechter presteren op de verbale tests dan hebben zij minder kans hebben om geselecteerd te worden voor een job en dan is er sprake van

discriminatie. Deze discriminatie is evenwel terecht als kennis van het Nederlands ook werkelijk essentieel is voor de succesvolle uitoefening van de job. Als het niet duidelijk is dat een betere kennis van het Nederlands leidt tot een betere jobperformantie dan is de discriminatie onterecht en dan zou men beter de verbale tests (waar de verschillen tussen Vlamingen en allochtonen het grootst zijn) niet gebruiken voor selectie. Merk op dat er bij SELOR en ABL in feite maar erg weinig eerste generatie allochtonen deelnemen aan selecties, zodat de discriminatie op basis van verbale tests in de praktijk zeer beperkt is.

Voor eerste generatie allochtonen hangt het presteren op verbale tests niet alleen af van algemene intelligentie, maar ook van de mate waarin ze reeds Nederlands hebben kunnen leren, wat ondermeer afhangt van verblijfsduur, leeftijd bij immigratie, enz. Verbale tests zouden daarom alleen mogen gebruikt worden voor selectie als de vaardigheden die ze meten essentieel zijn voor het succesvol uitoefenen van de job.

Voor selecties waar veel allochtonen aan deel nemen is het aangeraden om een taaltest af te nemen (zoals voorzien bij de MCT-M). Als de taalkennis van de kandidaat te laag is dan kan men het potentieel van de kandidaat beter bepalen door te vergelijken met allochtonen met een gelijkaardig profiel in termen van verblijfsduur en opleidingsniveau.

2. We stellen vast dat allochtonen niet alleen minder vaardig zijn bij "power tests", maar dat ze ook minder snel en minder nauwkeurig werken bij "snelheidstests". Bij het oplossen van de MCT-M laten allochtonen bijvoorbeeld meer items open hoewel er niet gecorrigeerd wordt voor raden bij het berekenen van testscores. Het is dus mogelijk dat allochtonen minder optimale antwoordstrategieën gebruiken en dat hun slechtere resultaten gedeeltelijk verklaard worden doordat ze minder "testslim" zijn. Om dit probleem te voorkomen volstaat het om alle kandidaten voor te bereiden op de test. Dit kan bijvoorbeeld door op voorhand uitleg te geven over mogelijke antwoordstrategieën bij het oplossen van meerkeuzevragen, oefenopgaven aan te bieden, en uit te leggen hoe definitieve testscores berekend worden (o.a. of er een correctie voor raden wordt toegepast).

Om er zeker van te zijn dat allochtonen en Vlamingen even vertrouwd zijn met optimale antwoordstrategieën bij het oplossen van tests, is het aangeraden om aan alle kandidaten op voorhand uitleg te geven over mogelijke antwoordstrategieën bij het oplossen van meerkeuzevragen, oefenopgaven aan te bieden, en uit te leggen hoe definitieve testscores berekend worden.

3. De resultaten van het onderzoek tonen aan dat alle onderzochte tests een aantal "discriminerende" DIF items bevatten die relatief moeilijker zijn voor Vlamingen of voor allochtonen met dezelfde totaalscore op de test. Bij de meeste tests zijn er evenwel slechts weinig items met DIF en zijn er in het algemeen slechts kleine verschillen tussen succeskansen van allochtonen en Vlamingen met gelijke vaardigheid. Uitzonderingen zijn de tests Woordrelaties, Woordanalogieën en Exclusie die in 55%, 50% en 37% van de items DIF vertonen. Vooral bij de verbale tests zijn er relatief veel items waar de DIF

op een beperkt deel van de schaal sterk oploopt (verschil in succeskansen groter dan .20). We kunnen concluderen dat verbale tests bij allochtonen en Vlamingen waarschijnlijk verschillende dimensies meten. Dit stemt overeen met de resultaten van factoranalyse op de subtestscores van allochtonen en Vlamingen. Bij allochtonen meten de verbale tests niet alleen algemene intelligentie maar ook schoolse kennis van Nederlands.

De resultaten van onze analyses tonen dat de DIF bij verbale tests gedeeltelijk kan verklaard worden. Items zijn bijvoorbeeld relatief moeilijker voor allochtonen als ze woorden bevatten die minder vaak voorkomen in het dagelijks taalgebruik, of als de woorden typisch zijn voor het Nederlands (de betekenis van de woorden kan minder gemakkelijk afgeleid worden op basis van een vreemde taal).

Het DIF onderzoek toont dat DIF vooral een probleem is bij de verbale tests. Bij gelijke vaardigheid hebben allochtonen en Vlamingen soms verschillende succeskansen. Items zijn bijvoorbeeld relatief moeilijker voor allochtonen als ze woorden bevatten die minder vaak voorkomen in het dagelijks taalgebruik, of als de woorden typisch zijn voor het Nederlands (de betekenis van de woorden kan minder gemakkelijk afgeleid worden op basis van een vreemde taal).

Dat er veel DIF is in verbale tests wil zeggen dat deze tests naast algemene intelligentie ook een andere dimensie meten zoals kennis van het Nederlands. De scores op de verbale tests kunnen dus niet gebruikt worden om Vlamingen en allochtonen te vergelijken op vlak van algemene intelligentie.

4. Het onderzoek toont aan dat voor alle onderzochte tests de gezamenlijke invloed van discriminatie in individuele items geen effect heeft op de testscores van Vlamingen en allochtonen. Hier moeten echter twee opmerkingen bij gemaakt worden. Ten eerste kunnen we stellen dat onze methode om het verschil in gemiddeld presteren te bepalen in de regel zal leiden tot positieve en negatieve DIF zodat op het niveau van de testscore meestal compensatie optreedt. Verder onderzoek is nodig om te evalueren in welke mate andere methoden om DIF te bepalen andere resultaten opleveren. Ten tweede, kunnen we stellen dat bij tests met veel DIF de testscores niet vergelijkbaar zijn omdat ze een verschillende betekenis kunnen hebben. Gelijke testscores kunnen namelijk het resultaat zijn van succes op verschillende verzamelingen van items die tegengestelde DIF vertonen. De constructvaliditeit van de test komt hierdoor in het gedrang.

We stellen ook vast dat bij vele tests de statistische onzekerheid in de testscore tamelijk groot is zodat het strikt ordenen van kandidaten met een klein verschil in scores niet zinvol is. Het is dan ook aangewezen om bij het ordenen van kandidaten rekening te houden met de onzekerheid in de testscores.

Voor de meeste tests kunnen testscores gebruikt worden om kandidaten op een valide manier te vergelijken. Voor tests met veel DIF (verbale tests) kunnen de testscores in principe niet vergeleken worden omdat ze een verschillende betekenis kunnen hebben.
Verder onderzoek is nodig om te evalueren in welke mate de methode om DIF te bepalen een invloed kan hebben op het bepalen van testbias.

Omdat de onzekerheid op de testscores soms erg groot is, zou het een standaardpraktijk moeten zijn om naast de testscore de onzekerheid te rapporteren zodat onbelangrijke verschillen tussen testscores niet kunnen doorwegen op de besluitvorming. Methoden uit de itemresponstheorie kunnen ook helpen om de vaardigheid van een persoon nauwkeuriger te schatten. Omdat de tests op zulke grote schaal gebruikt worden zou men hier zeker alle inspanningen moeten doen om personen optimaal te ordenen.

5. Om een gelijkekansenbeleid te voeren dient men te vermijden dat kansengroepen ten onrechte gediscrimineerd worden op basis van tests. Dit kan door de kwaliteit van tests op een aantal punten systematisch te bewaken. Zoals bij geneesmiddelen en voedingswaren al het geval is zou men de kwaliteitscontrole op tests die gebruikt worden in selecties bij de overheid en in de privéondernemingen kunnen verscherpen en systematiseren. Als een nieuwe test op de markt gebracht wordt door een testproducent zou men de test een kwaliteitslabel kunnen toekennen dat aangeeft in welke mate de test onderzocht is op aspecten als betrouwbaarheid, validiteit, afwezigheid van discriminatie ten aanzien van kansengroepen enz. Het toekennen van een kwaliteitslabel zou bijvoorbeeld kunnen gebeuren door een onafhankelijke commissie van testexperts (zoals bijvoorbeeld de Commissie Test Aangelegenheden Nederland die is ingesteld door het Nederlands Instituut van Psychologen).

6. Om in de toekomst de kwaliteit van tests (bij de overheid) te bewaken en te optimaliseren dient men de ruwe testgegevens en relevante achtergrondvariabelen na de selectie te inventariseren. Vervolgens kan men op basis van de analyse van deze gegevens de kwaliteit van de tests optimaliseren. Deze optimalisatie zou bijvoorbeeld kunnen gebeuren door testexperts in dienst van de overheid, of zou bijvoorbeeld kunnen uitbesteed worden aan een wetenschappelijke instelling.

7. Hoe een goede en betrouwbare rangordening te maken van kandidaten op basis van de scores die behaald werden op basis van één of meerdere selectietests blijkt een belangrijk vraagstuk te zijn dat ook ter sprake komt in het eerste deelrapport bij redelijke aanpassingen voor personen met een handicap. In de besprekingen in verband met dit onderzoek blijkt dat een afstemming op het terrein tussen enerzijds de verwachtingen van de overheid als opdrachtgever en anderzijds de selecteurs, rekening houdende met de aanbevelingen in de deelrapporten, zich opdringt.

Referenties

Born, M. P., Bleichrodt, N. & Van der Flier, H. (1987). Cross-cultural comparison of sex-related differences on intelligence tests. *Journal of Cross-Cultural Psychology, 18,* 283-314.

De Jong, M. J., & van Batenburg, T. A. (1984). Etnische herkomst, intelligentie en schoolkeuzeadvies. *Pedagogische studiën, 61,* 362-371.

Evers, A., & Lucassen, W. (1991). *Handleiding DAT*. Lisse: Swets & Zeitlinger.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Hines, M. (1990). Gonadal hormones and human cognitive development. In J. Balthazart (ed), *Brain and behavior in vertebrates 1: sexual differentiation neuroanatomical aspects, neurotransmitters and neuropeptides*, 51-63. Basel, Zwitserland: Kruger

Kaufman, A. S. & Horn, J. L. (1996). Age changes on tests of Fluid and Crystallized ability for women and men on the Kaufman Adolescent and Adult Intelligence Test (KAIT) at ages 17-94 years. *Archives of Clinical Neuropsychology, 11,* 97-121.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Pieters, J. P. M., & Zaal, J. N. (1991). Culturele bias in de Nederlandse Politie Intelligentie Test: waar psychologie eindigt en beleid begint. In H. van der Flier, P. G. W. Jansen & J. N. Zaal (eds), *Selectieresearch in de praktijk,* 247-261. Lisse: Swets & Zeitlinger

Poortinga, Y. H., & Flier, H. van der (1988). The meaning of item bias in ability tests. In Irvine, S. H. & Berry, J. W. (eds), *Human abilities in cultural context,* Cambridge: Cambridge University Press.

Resing, W. C. M., Bleichrodt, N. & Drenth, P. J. D. (1986). Het gebruik van de RAKIT bij allochtoon etnische groepen. *Nederlands Tijdschrift voor de Psychologie, 41,* 179-188.

Schmitt, A. P. & Dorans, N. J. (1990). *Differential item functioning for minority examines on the SAT.* Journal of Educational Measurement, 27, 67-81.

Shealy, R. T., & Stout, W. F. (1993a). An item response theory model for test bias and differential test functioning. In P. W. Holland and H. Wainer (eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R. T., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF, *Psychometrika, 58*, 159-194.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). WinBUGS version 1.2 user manual. Cambridge, UK: MRC Biostatistics Unit.

Te Nijenhuis, J. (1997). *Comparability of test scores for immigrants and majority group members in the Netherlands.* Academisch proefschrift, Vrije Universiteit.

Te Nijenhuis, J. & Van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: some Dutch findings. *Journal of Applied Psychology, 82*, 675-687.

Van den Berg, R. H. (2001). *Psychologisch onderzoek in een multiculturele samenleving: Psychologische tests, interview- en functioneringsbeoordelingen*. Academisch proefschrift, Vrije Universiteit.

Van Leest, P. F., & Bleichrodt, N. (1990). Testing of collegegraduates from ethnic minority groups. In N. Bleichrodt & P. J. D. Drenth (eds), *Contemporary issues in cross-cultural psychology: Selected papers from the regional conference of the International Association of Cross-Cultural Psychology*. Amsterdam: Swets & Zeitlinger.

Zimowski, F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1994). *BIMAIN 2: Multiple group IRT analysis an test maintenance for binary items*. Chicago: Scientific Software International.

Bijlage 1: Vragenlijst naar achtergrondgegevens

Beste sollicitant,

De Vlaamse Regering vindt het erg belangrijk dat iedereen dezelfde kansen krijgt op de arbeidsmarkt. Daarom heeft de overheid aan de KU Leuven de opdracht gegeven om te onderzoeken of intelligentietests die gebruikt worden in personeelsselectie eerlijk zijn voor allochtonen. Voor dit onderzoek vragen we aan sollicitanten om een korte vragenlijst in te vullen. Wij zouden het erg op prijs stellen indien u zou willen meewerken aan dit onderzoek.

We willen er de nadruk op leggen dat de wet op de Privacy (Koninklijk Besluit, 8/12/1992) van toepassing is op de vragenlijst. Alle informatie zal uitsluitend in het kader van het onderzoek gebruikt worden en dus niet worden doorgegeven aan de beoordelaars van uw sollicitatie. Om de anonimiteit van de gegevens te garanderen vragen wij u om de ingevulde vragenlijst in de bijgeleverde enveloppe te steken, deze zorgvuldig toe te plakken en in de daarvoor voorziene doos te stoppen.

Alvast hartelijk bedankt voor uw medewerking en nog veel succes met uw sollicitatie!

Vriendelijke groeten

Het onderzoeksteam van de KU Leuven:

Michel Meulders
Paul De Boeck
Karel De Witte
Rianne Janssen
Miek Vandenberk

Vul hier de **laatste 6 cijfers van uw rijksregisternummer** in. Deze code kan op geen enkele manier gebruikt worden om uw identiteit te achterhalen.

[Ter informatie: Het rijksregisternummer vindt u op de achterzijde van uw identiteitskaart of op de voorkant van uw SISkaart. De eerste 6 cijfers van dit nummer bestaan uit uw geboortedatum.]

In welk land bent u geboren?
……...

Welke nationaliteit had u bij uw geboorte?
……...

Hoe oud bent u?
……...

Hoe lang bent u al in België (in jaren)?
……..

Wat is uw moedertaal?
……..

Welke taal spreekt u thuis het meest?
……...

Met welke bevolkingsgroep voelt u zich het meest verbonden? (Walen, Vlamingen, Koerden, …)
……...

Indien uw vader, moeder, en grootouders de Belgische nationaliteit hadden bij hun geboorte moet u de volgende vragen niet beantwoorden. Indien dit niet het geval is, gelieve dan de volgende vragen zo nauwkeurig mogelijk in te vullen.

Welke nationaliteit had uw vader bij zijn geboorte?
……...

Welke nationaliteit had uw moeder bij haar geboorte?
……...

Welke nationaliteit had uw grootvader langs vaders zijde bij zijn geboorte?
……...

Welke nationaliteit had uw grootmoeder langs vaders zijde bij haar geboorte?
……...

Welke nationaliteit had uw grootvader langs moeders zijde bij zijn geboorte?
……...

Welke nationaliteit had uw grootmoeder langs moeders zijde bij haar geboorte?
……...

Nogmaals hartelijk dank voor uw medewerking!

Bijlage 2: Verkorte beoordelingsschaal van woorden van verbale tests

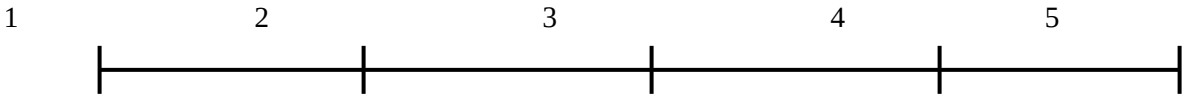
Onderzoek Verbale tests MCT-M

Naam:	**E-mail:**
Niveau van de cursisten waaraan u Nederlands geeft:	

BESCHRIJVING ONDERZOEK

In de lijst op de volgende bladzijde staan een aantal woorden.
We zouden graag het volgende weten:

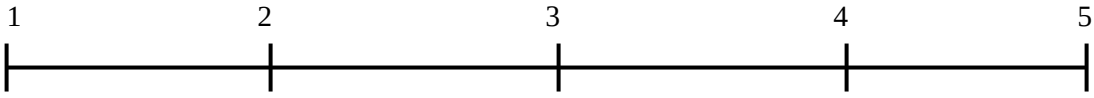
1) Hoe vertrouwd zijn allochtonen met deze woorden?
U kunt volgende schaal gebruiken om uit te drukken hoe dikwijls zij (waarschijnlijk) te maken krijgen met een bepaald woord:



Nooit **Bijna nooit** **Soms** **Dikwijls** **Heel dikwijls**

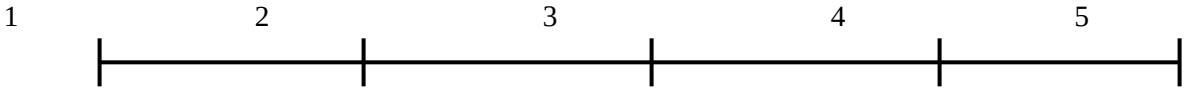
1. **NOOIT** (Ze hebben het woord waarschijnlijk nog *nooit* gezien, gehoord of gebruikt.)
2. **BIJNA NOOIT** (Ze hebt dit woord ten minste al één keer gezien, gehoord of gebruikt, maar daarna bijna *nooit meer*.)
3. **SOMS** (Ze zien, horen of gebruiken dit woord *soms*, maar niet dikwijls.)
4. **DIKWIJLS** (Ze zien, horen of gebruikt het woord wel *dikwijls*, maar niet heel dikwijls.)
5. **HEEL DIKWIJLS** (Ze zien, horen of gebruiken dit woord *bijna elke dag*.)

2) Hoe concreet vindt u het woord?



Heel concreet **Heel abstract**

3) Kan men dit woord gemakkelijk kennen op basis van een vreemde taal (Engels, Frans, …)?
Vb: Tactiek: in het Engels 'tactic', dus heel gemakkelijk te kennen o.b.v. het Engels
Professor: in het Frans 'professeur', dus heel gemakkelijk te kennen o.b.v. het Frans



Heel gemakkelijk **Heel moeilijk**

Bedankt voor uw medewerking en veel succes!

	Hoe dikwijls krijgen allochtonen te ma het woord?					Hoe concreet vindt u het					Hoe makkelijk kan men dit woord o.b.v. een andere taal?				
	Nooit	Bijna nooit	Soms	Dikwijls	Heel dikwijls	Heel concreet				Heel abstract	Heel makkelijk				Heel moeilijk
blind	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
oog	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
doof	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
gehoor	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
woord	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
oor	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
stem	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
geluid	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
winkelier	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
verkopen	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
klant	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
winst	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
kopen	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
product	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
rekening	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
winkel	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
dood	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
geboorte	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
verlies	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
jong	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
leven	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
winst	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
oud	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
ziek	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
maand	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
jaar	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
dag	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
nacht	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
week	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
tijd	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
eeuw	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
uur	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
hand	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
handschoen	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
voet	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
teen	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
sok	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
broek	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
vinger	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
been	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5