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 mannen en vrouwen gelijke kansen bij selectieproeven met intelligentietests?

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

Mei 2004

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 vrouwen 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 60 67
E-mail: miek.vandenberk@psy.kuleuven.ac.be

Hoofdstuk 1: Inleiding

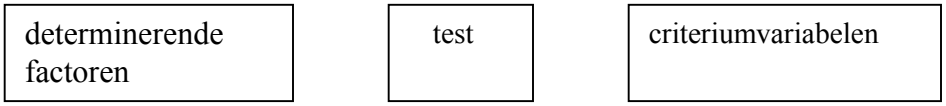
De bedoeling van het project is om te onderzoeken of relevante intelligentietests die gebruikt worden in de context van personeelsselectie een bias vertonen voor elk van drie kansengroepen, namelijk vrouwen, allochtonen en gehandicapten. In dit rapport beschrijven we de resultaten van het onderzoek bij de doelgroep vrouwen. Hiervoor baseren we ons op de analyse van testgegevens die verzameld werden bij het 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 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 Katrien Brysse (SELOR) en Bert Schreurs (ABL) die steeds zo vriendelijk waren om ons verder te helpen.

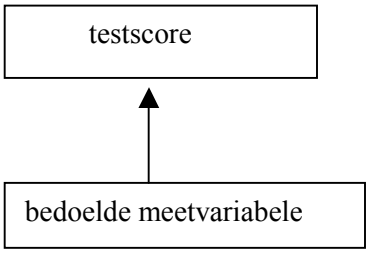
1.1 Conceptueel kader

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.

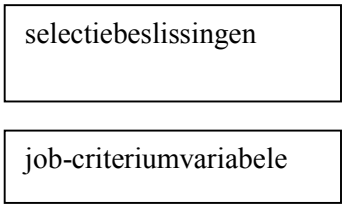
b. Een test is altijd bedoeld om een bepaalde persoonskarakteristiek te meten: de variabele die de test bedoelt te meten, verder ook de *bedoelde meetvariabele* genoemd. Die karakteristiek hoeft geen onveranderlijke karakteristiek te zijn, het kan ook gaan om een niveau dat men tijdelijk heeft bereikt 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.



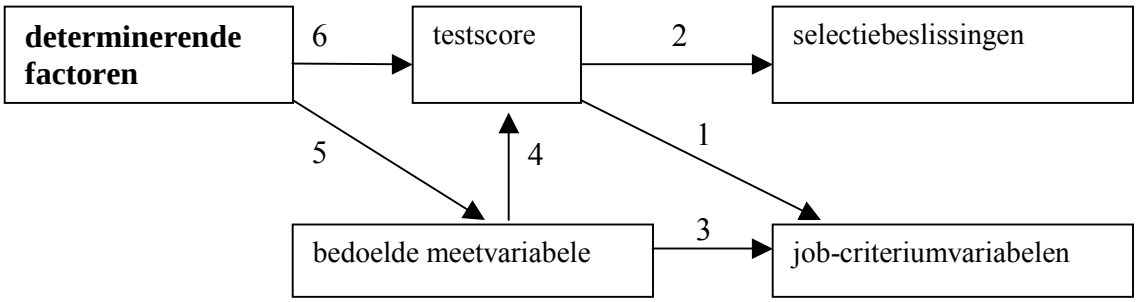
Men kan de band met de bedoelde meetvariabele per item bekijken. In principe speelt de bedoelde meetvariabele een rol in elk item, maar naargelang van het item kan die variabele sterker of minder sterk doorwegen. Het gewicht van de bedoelde meetvariabele in een item noemt men de "discriminatiegraad"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

a. Bias-onderzoek 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 bias-onderzoek (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. Bias-onderzoek 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 bias-onderzoek. 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 bias-onderzoek. 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 Onderzoeksplan

Het onderzoek naar bias bij de doelgroep vrouwen bestaat uit vier fasen. In de eerste fase worden testgegevens opgevraagd 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 anwoorden dat gecorrigeerd is voor raden bij meerkeuze-vragen) 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 vier fasen worden beschreven in de Hoofdstukken 2 t.e.m. 5: Hoofdstuk 2 bevat een beschrijving van de criteria die gehanteerd werden bij het selecteren van de tests alsook een overzicht van de opgevraagde datasets. Hoofdstuk 3 bevat een algemeen theoretisch overzicht van de IRT aanpak voor biasonderzoek. In hoofdstuk 4 wordt per test het resultaat van de fasen 2 tot 4 beschreven. Hoofdstuk 5 bevat een synthese van de wetenschappelijke bevindingen en de beleidsaanbevelingen die hieraan kunnen gekoppeld worden.

Hoofdstuk 2: Dataverzameling

2.1 Selectie van intelligentietests

De bedoeling van dit rapport is te onderzoeken of intelligentietests die veel gebruikt worden voor personeelsselectie discriminerend zijn voor vrouwen versus mannen. Zowel tests die gebruikt worden voor personeelsselectie bij de overheid als bij privé-ondernemingen zijn hierbij van belang. Voor de selectie van overheidspersoneel werd samenwerking gezocht met *SELOR.* De testbatterij van SELOR bestaat uit 14 computertests (zie Tabel 2.1). Hieruit werden vier tests gekozen:

- Logded: logische deducties
- Anaverb: verbale analogieën
- Codes: code leren
- Numva: cijferreeksen

Bij het kiezen van deze tests is met verschillende criteria rekening gehouden: (1) de frequentie waarmee de test wordt afgenomen dient voldoende hoog te zijn, (2) de tests dienen verschillende soorten intelligentie te meten, (3) er dienen voldoende gegevens voorhanden te zijn om het biasonderzoek op een betrouwbare manier uit te kunnen voeren, (4) om de resultaten van het biasonderzoek te kunnen verklaren is het van belang dat de testitems kunnen ontleed worden in termen van de cognitieve processen die nodig om de test op te lossen.

Daarnaast werd ook contact opgenomen met Defensie (verder afgekort als ABL*)*. Een eerste stap van alle kandidaten die in de provinciale defensiehuizen informeren voor een functie bij ABL is het invullen van de PinP. Dit is een testbatterij die bestaat uit 3 computertests:

- DGEO: spatiale oriëntatie, ruimtelijk voorstellingsvermogen
- TNV: algemene intelligentie
- WIMA: numerieke vaardigheden

Deze tests beantwoorden aan de hoger vermelde criteria Aangezien dit project uitgaat van de Vlaamse regering beperken we ons biasonderzoek tot de gegevens die verzameld zijn in defensiehuizen in Vlaanderen en Brussel.

Drie van de vier tests van SELOR (Anaverb, Codes, Logded) werden ontwikkeld door het bedrijf CEBIR dat computertests ontwikkelt om intelligentie en persoonlijkheid te meten. De NUMVA test werd ontwikkeld door SELOR zelf. Aangezien de door CEBIR ontwikkelde tests veel gebruikt worden voor personeelsselectie in privé-ondernemingen zijn ze zeer relevant voor deze context**.** Er werden daarom verder geen gegevens voor privé-ondernemingen opgevraagd.

Tabel 2.1. Computertests van SELOR

Test	**Soort intelligentie**	**Frequentie afname Nederlands- en Frans-taligen ('97-'02)**		**cognitieve analyse?**	**Opmerking**
		mannen	**vrouwen**		
Admin	Perceptuele Snelheid Geïntegreerde processen	2613	1711	-	
Alphabet	Inductie	8797	7835	+	Adaptief
Anaverb (verbale analogieën)	Inductie Verbaal Begrip	7772	7765	+	
Arianew	Verbaal Begrip	16200	22151	?	Ingewikkeld
Codes (code leren)	Inductie Geïntegreerde processen	4648	4038	+	Dynamisch
Logded (transitief redeneren)	Deductie	9580	9322	+	
Logix	Deductie	3124	2416	+	
Numap	Getalbegrip	2604	2735	?	
Numva (cijferreeksen)	Inductie	1902	1938	+	
Percep	Perceptuele Snelheid Geheugenspan	3792	2646	-	
Metrolog	Redeneren	48	30	?	
Problems	Redeneren	373	169	?	
Pairnumb	Perceptuele Snelheid	318	324	-	
Diag (stroomdiagrammen)	Deductie	593	480	?	Specifiek

Opmerking: Alle tests zijn snelheidstests met een tijdslimiet, dus niet alle items worden door alle deelnemers beantwoord.

2.2 Beschrijving tests

2.2.1 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, etc. 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 meerkeuze-vragen met elk 5 antwoordalternatieven. De kandidaten krijgen 15 minuten de tijd om de test op te lossen. Per test tonen we hier een kader met een voorbeeld item. Het juiste antwoord is telkens 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

2.2.2 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 meerkeuze-vragen met elk 4 antwoordalternatieven. De kandidaten beschikken over 20 minuten om de test op te lossen.

Voorbeelditem:

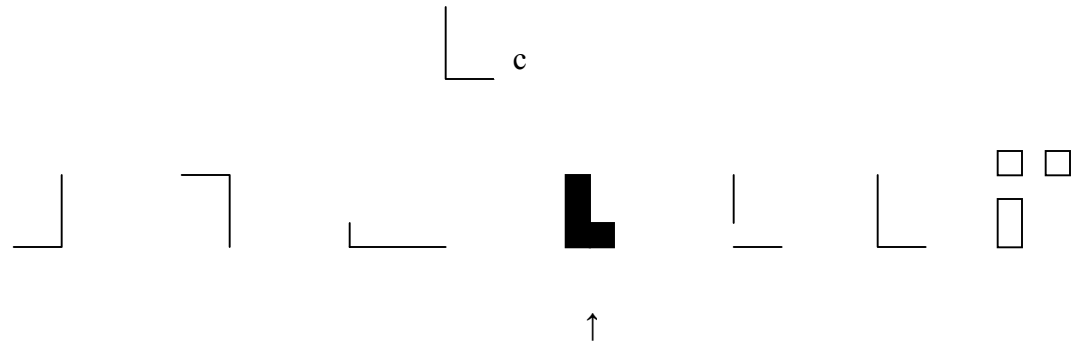
Lood	pluim		Veder
Zwaar	?		Mooier
			Lucht
		→	Licht

2.2.3 CODES

Met de CODE test 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:



2.2.4 NUMVA

De NUMVA test meet 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 dient op te lossen in 30 minuten.

Voorbeelditem:

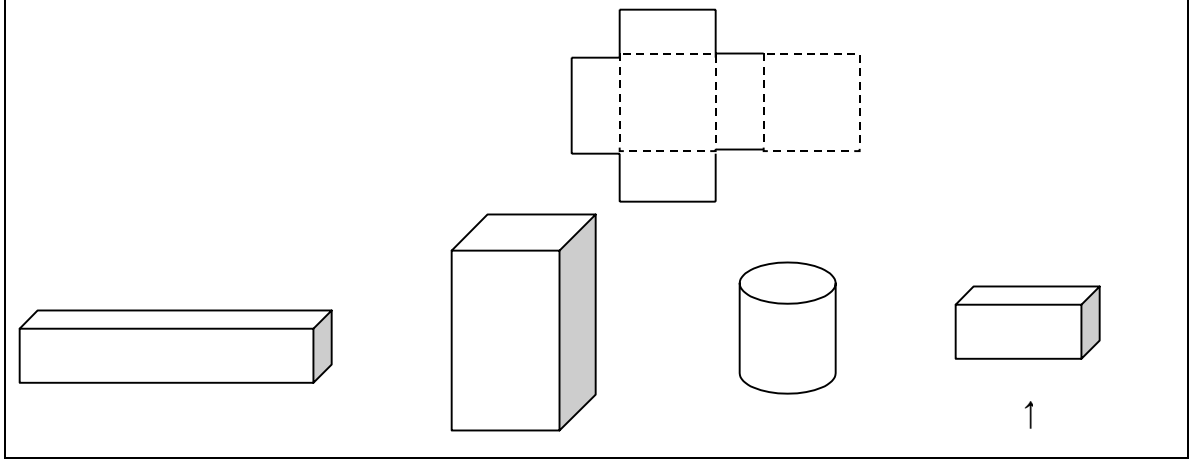
4 6 9 13 18 24 …

30
6
→ 31
12

2.2.5 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 dienen opgelost te worden in 7 minuten.

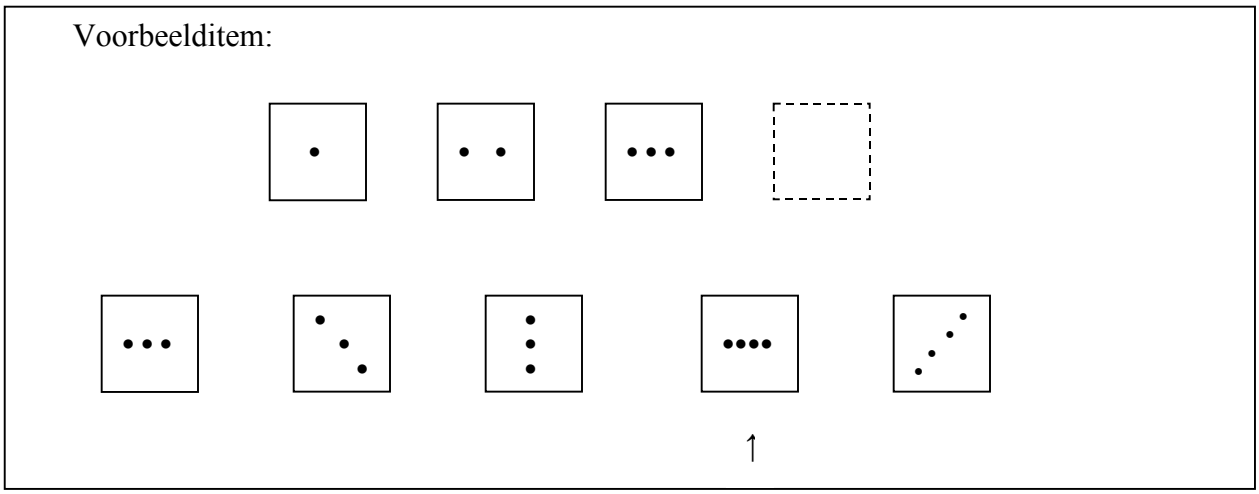
Voorbeelditem:



2.2.6 TNV

Met de TNV test meet men de 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 meerkeuze-vragen met elk 5 antwoordalternatieven en moet opgelost worden in 15 minuten.

Voorbeelditem:



2.2.7 WIMA

Met de WIMA test meet men de vaardigheid om numerieke vraagstukken op te lossen De test bevat 23 meerkeuze-vragen met telkens 4 antwoordalternatieven. De kandidaten krijgen 30 minuten om de test op te lossen. Bij het oplossen van de test beschikt men over 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

2.3 Beschrijving van opgevraagde testgegevens

Voor bovenstaande tests werden bij SELOR en ABL de ruwe testgegevens opgevraagd van mannen en vrouwen die in de afgelopen jaren aan een selectie deelnamen. Tabel 2.2 geeft per test een overzicht van enkele karakteristieken van de gegevens: 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. Alle datasets bevatten voldoende gegevens om een betrouwbare biasanalyse te kunnen uit voeren.

Tabel 2.2 Karakteristieken van de onderzochte datasets

TEST	aantal vb-items	aantal items	aantal antw-altern.	tijds-limiet	correctie voor raden	groep	aantal personen	Cronbach α
						mannen	1895	0,80
LOGDED	2	22	5	15 min	ja	vrouwen	2521	0,73
						mannen	1656	0,92
ANAVERB	2	100	4	20 min	ja	vrouwen	2401	0,90
						mannen	731	0,96
CODES	2	74	7	45 min	nee	vrouwen	574	0,95
						mannen	1113	0,84
NUMVA	2	38	4	30 min	ja	vrouwen	1227	0,80
						mannen	2962	0,87
DGEO	5	40	4	7 min	ja	vrouwen	431	0,86
						mannen	2963	0,87
TNV	5	50	5	15 min	ja	vrouwen	431	0,89
						mannen	2257	0,89
WIMA	2	23	4	30 min	ja	vrouwen	341	0,88

Tabel 2.2 laat zien dat de meeste tests een hoge tot zeer hoge betrouwbaarheid hebben (α hoger dan .85). Uitzonderingen zijn de korte test LOGDED en de NUMVA test. Verder stellen we vast dat de betrouwbaarheid van de meeste tests ongeveer even hoog is voor mannen als voor vrouwen. Uitzonderingen zijn LOGDED en ANAVERB die telkens een iets hogere betrouwbaarheid hebben voor de groep mannen.

2.4 Scoringsvoorschriften

De bovenvermelde datasets bestaan uit een persoon-bij-itemmatrix gevuld met nullen en énen waarbij elk element aangeeft of een bepaalde persoon een bepaald item juist (score 1) of fout (score 0) heeft opgelost. Items die werden opengelaten wegens tijdsgebrek of omdat ze te moeilijk waren worden ook fout gerekend.

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

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

Als illustratie van de logica achter deze gecorrigeerde totaalscore beschouwen we het volgende voorbeeld.

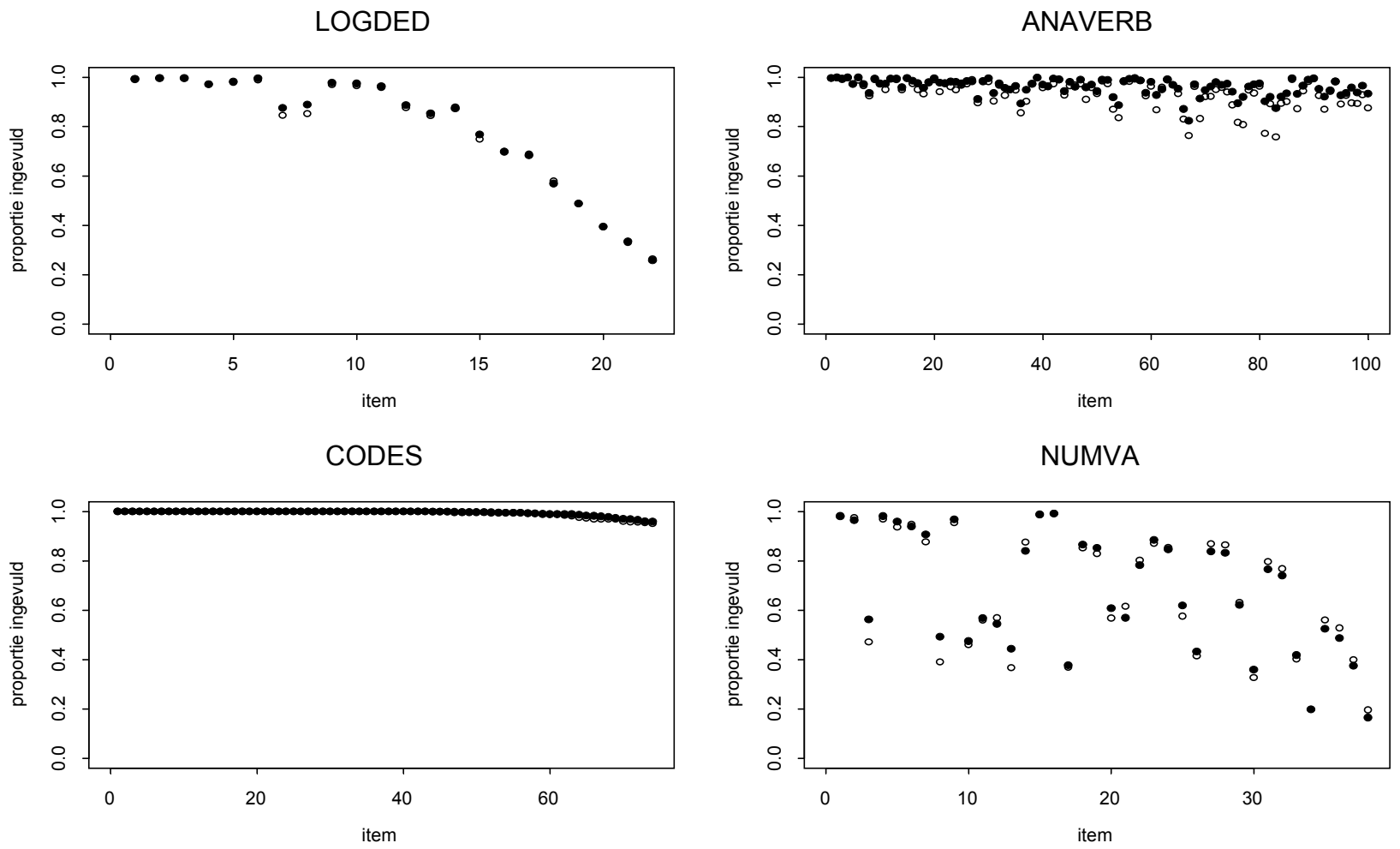
Stel: Kandidaat A raadt bij 12 items met 4 antwoordalternatieven. Op basis van toeval zou deze kandidaat 3 punten extra krijgen (0.25 *12). Wanneer men de formule van de gecorrigeerde totaalscore uitrekent, trekt men bij het aantal juiste antwoorden 3 punten af, zodat er gecorrigeerd wordt voor raden ((9/(4-1)) = 3). Dus men gaat zo streng straffen als wat de kandidaat gemiddeld op toeval wint door te raden.

We merken nog op dat voor alle tests een strikte tijdslimiet geldt. Dit heeft als gevolg dat de items op het einde van een test soms niet meer kunnen ingevuld worden wegens tijdsgebrek. Ter illustratie toont Figuur 2.1 voor elke test de proportie mannen en vrouwen die een bepaald item invullen.

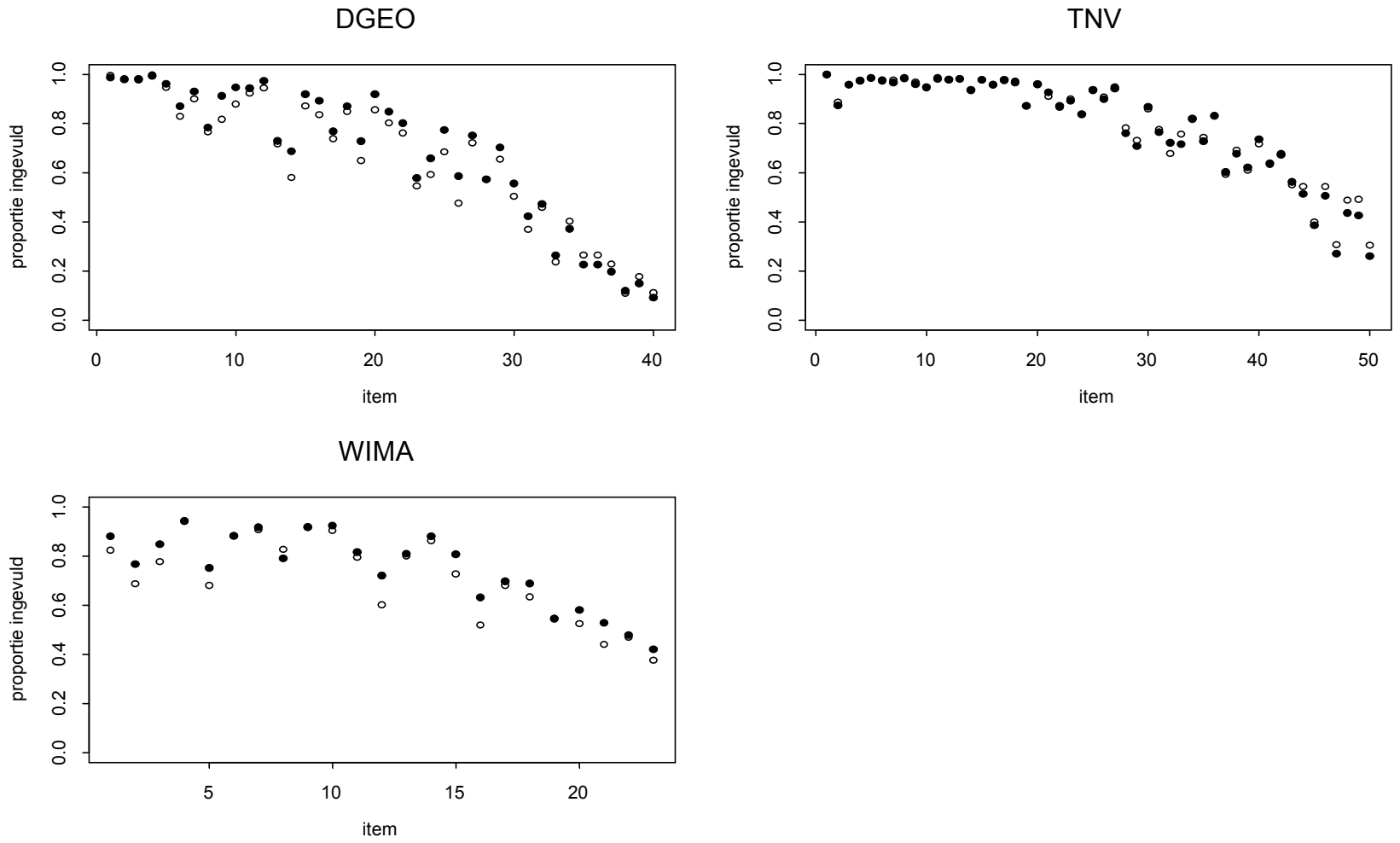
Een eerste observatie is dat het effect van de tijdslimiet op het niet invullen van items sterk kan verschillen van test tot test: Bij sommige tests is de tijdslimiet niet erg streng zodat de meeste items door bijna alle kandidaten ingevuld worden (ANAVERRB en CODES). Bij andere tests is de tijdslimiet wel redelijk streng zodat de proportie kandidaten die een item invullen sterk daalt naar het einde van de test toe (LOGDED,

DGEO, TNV, WIMA). Tot slot stellen we vast dat bij NUMVA bepaalde (waarschijnlijk moeilijke) items door relatief veel kandidaten niet ingevuld worden.

Een tweede observatie is dat het verschil tussen mannen en vrouwen bij het niet invullen van items ook varieert van test tot test: Bij LOGDED en CODES is de proportie mannen en vrouwen die een item invullen ongeveer gelijk. Bij de andere tests treden er over het algemeen wel verschillen op tussen de twee groepen. Merk op dat deze verschillen een mogelijke bron van item bias kunnen zijn. Om na te gaan of dit effectief het geval is zou men kunnen onderzoeken of het verband tussen de onderliggende vaardigheid en het niet invullen van items verschilt in de twee groepen.



Figuur 2.1: Proportie mannen (●) versus vrouwen (o) die een bepaald item invullen



Vervolg Figuur 2.1: Proportie mannen (●) versus vrouwen (o) die een bepaald item invullen

Hoofdstuk 3: Een IRT Benadering voor biasonderzoek

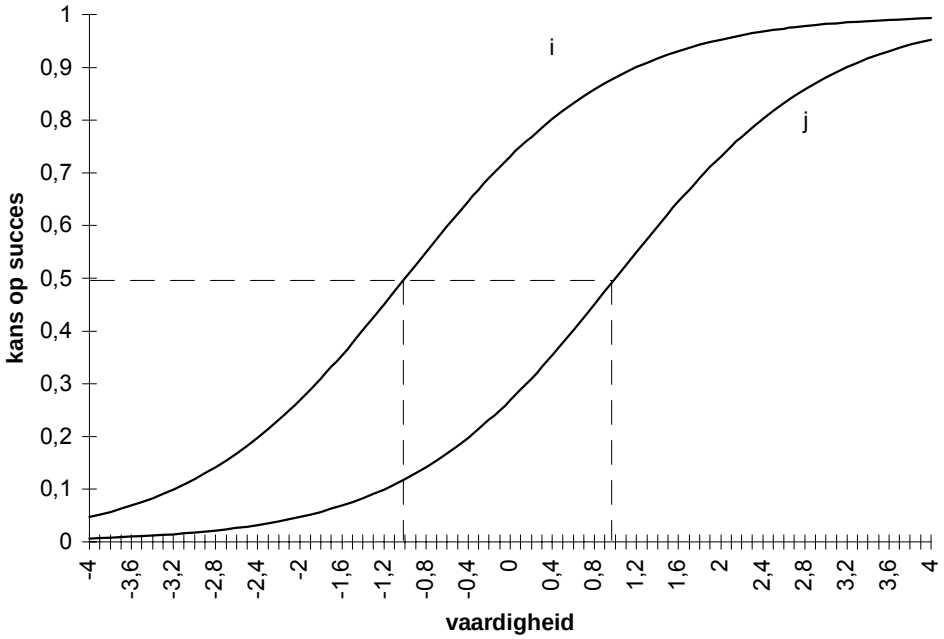
3.1 Itemresponsmodellen

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:

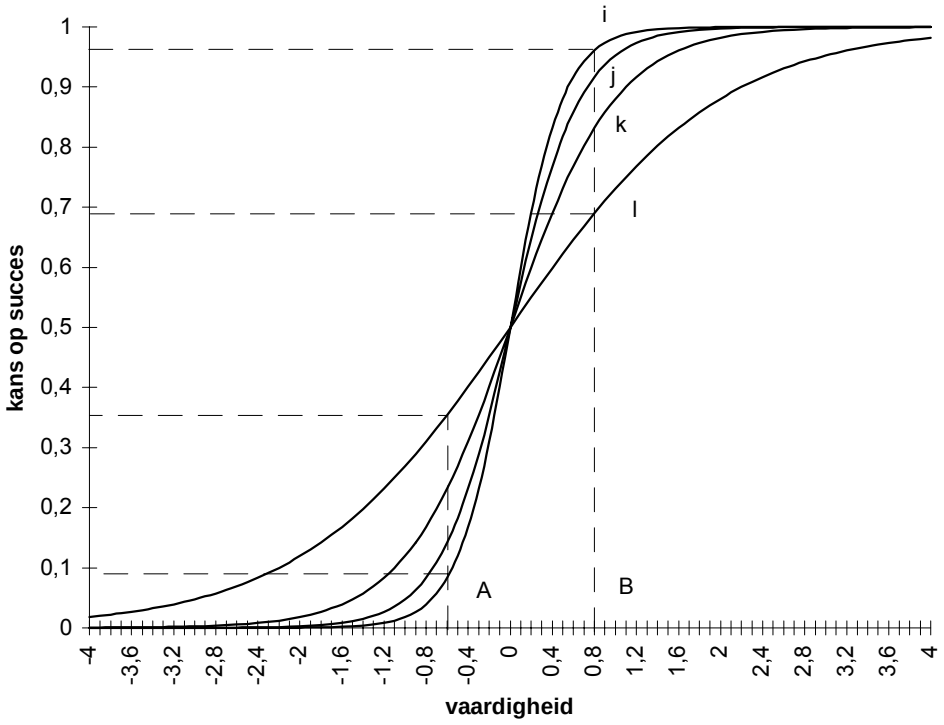
$$\Pr(Y_{pi} = 1 \mid \theta_p, \alpha_i, \beta_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{\exp\left[\alpha_i(\theta_p - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_p - \beta_i)\right]} \qquad (3.1)$$

Hierbij verwijst θ_p 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 α_i, β_i en γ_i hebben een specifieke interpretatie.

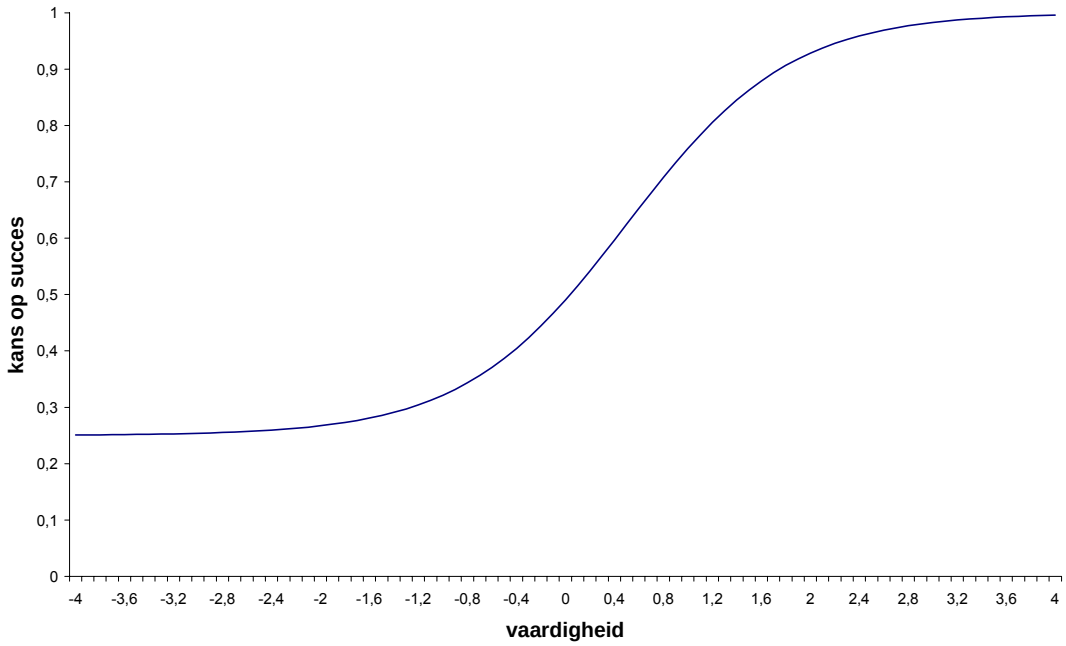
De parameter γ_i, ook wel raadparameter genoemd, is de kans om het item juist op te lossen als men een oneindig lage vaardigheid heeft. In de grafische voorstelling is deze parameter de linkerasymptoot van de itemresponsfunctie. Bij open vragen is het redelijk om deze parameter op voorhand gelijk te 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 β_i 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 ($\alpha_i = \alpha_j = 1$) en verschillende moeilijkheidsgraad $\beta_i = -1$ en $\beta_j = 1$.



Figuur 3.2 Itemresponsfuncties voor items met gelijke moeilijkheidsgraad $\beta_i = \beta_j = \beta_k = \beta_l = 0$ en verschillende discriminatiegraad $\alpha_i = 4$, $\alpha_j = 3$, $\alpha_k = 2$ en $\alpha_l = 1$.



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 δ_i 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 vrouwen, Z=1 voor mannen), 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.

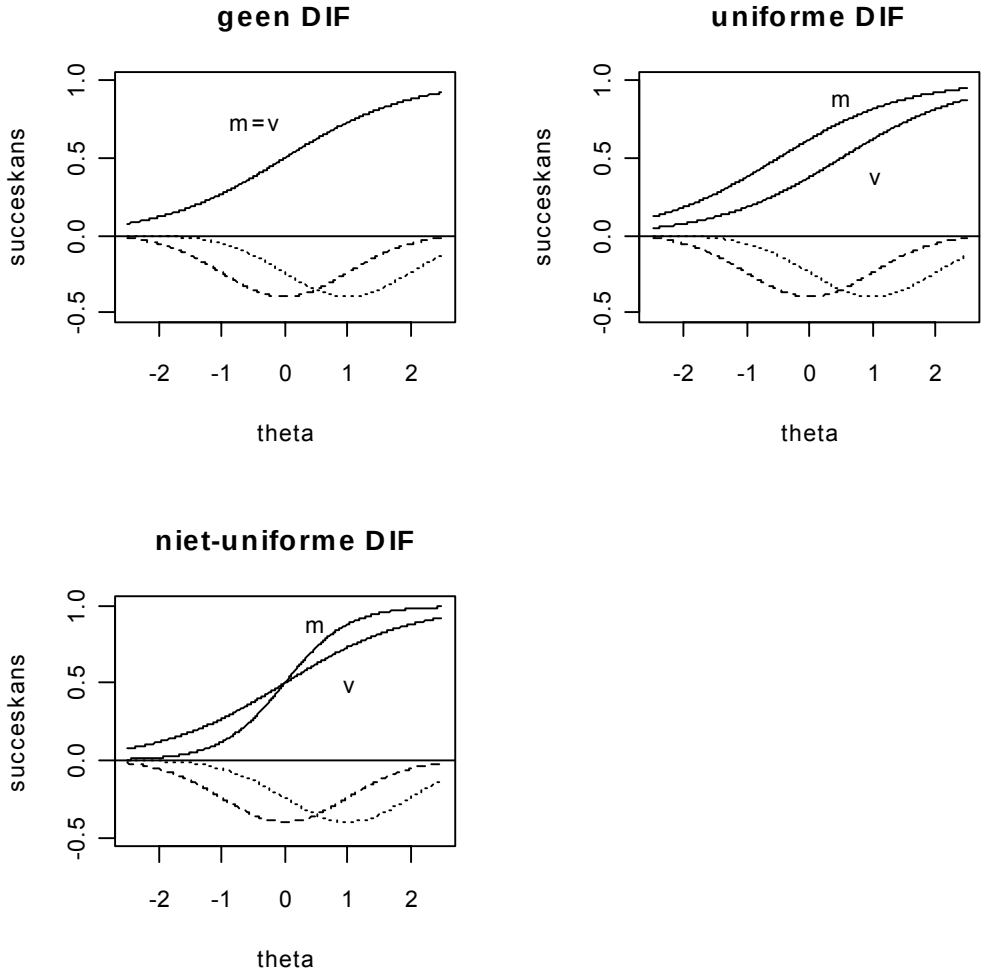
Het linker-boven paneel in Figuur 3.4 toont de itemresponsfunctie van een item dat geen DIF vertoont. We merken op dat beide groepen wel een andere vaardigheidsverdeling kunnen hebben. In Figuur 3.4 hebben vrouwen gemiddeld een hoger vaardigheidsniveau dan mannen en is de spreiding van de vaardigheid 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 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 moeilijkheidsgraden of 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 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 mannen (m) en vrouwen (v) voor een item zonder DIF, met uniforme DIF en met niet-uniforme DIF. De vaardigheidsverdelingen voor mannen (- - -) en vrouwen (….) 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:

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

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

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

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 vrouwen (Z=0) gelden moeilijkheidsgraden β_i en discriminatiegraden α_i en voor mannen (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.

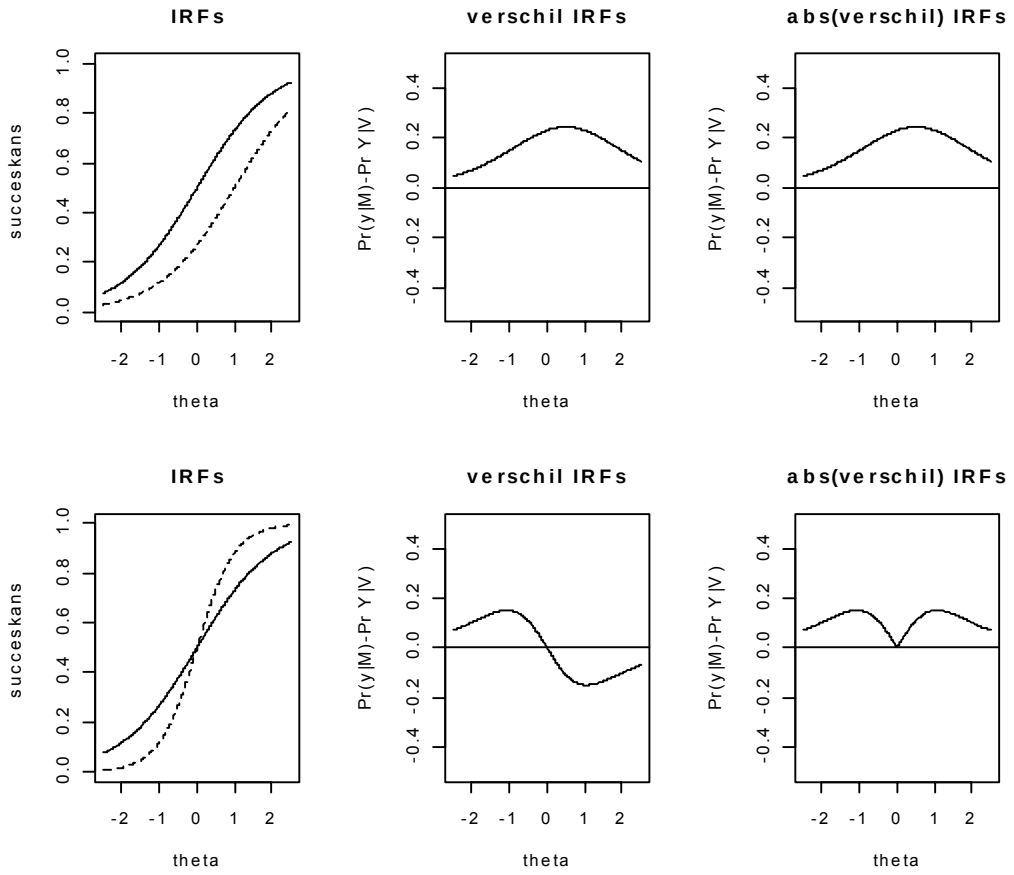
In formules (3.2)-(3.3) veronderstellen we ook dat de latente variabele θ een verschillende verdeling kan hebben naargelang van de groep, namelijk $\theta \sim N(0,1)$ voor vrouwen en $\theta \sim N(\mu,\sigma^2)$ voor mannen. We merken hierbij op dat de parameters van de normale verdeling bij vrouwen 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 aan van de gemiddelde man op de schaal terwijl de positie van de gemiddelde vrouw 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 mannen en vrouwen 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 $\xi_i\neq 0$). 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 ($\varepsilon_i\neq 0$) 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 wordt zoveel mogelijk constant gehouden voor alle datasets die besproken worden in dit rapport. Deze strategie bestaat uit drie stappen:

(1) De parameters van het 3PL worden geschat voor elke groep apart en ze worden op een gemeenschappelijke schaal geplaatst op basis van de methode van gelijke populatiegemiddelden.

(2) Er wordt aan de hand van statistische tests voor elk item nagegaan of er DIF optreedt in de moeilijkheidsgraad ($\xi_i \neq 0$) of in de discriminatiegraad ($\varepsilon_i \neq 0$). Als voor slechts enkele items van de

ganse test de discriminatiegraden verschillen tussen groepen, dan wordt statistisch getoetst of een model met alleen uniforme DIF houdbaar is. Als dit het geval is wordt verder gewerkt met een model dat alleen uniforme DIF toelaat. Als dit niet het geval is wordt verder gewerkt met het model dat niet-uniforme DIF toelaat voor elk item.

(3) Indien mogelijk wordt er een set van anker-items gekozen waarvoor het verloop van de itemresponsfuncties zeer sterk gelijkend is in de twee groepen (ξ_i en ε_i ongeveer gelijk aan 0).
Een analyse op de twee groepen tegelijk met dit anker als vergelijkingsbasis geeft dan in principe hetzelfde resultaat als op basis van de methode van gelijke populatiegemiddelden. Om het gekozen anker te valideren wordt deze controle ook effectief uitgevoerd. Voor een aantal data sets (in dit rapport de data sets waar alleen uniforme DIF optreedt) wordt het anker dan verder gebruikt bij analyses om DIF te verklaren op basis van itemkenmerken. Als het niet mogelijk was om een goed anker te kiezen dan wordt voor het verklaren van DIF verder gewerkt met de DIF resultaten die bekomen werden op basis van de methode van gelijke populatiegemiddelden.

Om de verschillende stappen van de analyse strategie praktisch uit te voeren gebruiken we verschillende software programma's. Om het 3PL te schatten voor één welbepaalde groep (eerste stap) maken we gebruik van het programma BIMAIN (Zimowski, Muraki, Mislevy & Bock, 1994) of van een zelf geïmplementeerd algoritme uit Bayesiaanse data analyse, namelijk het Metropolis algoritme(Gelman, Carlin, Stern & Rubin, 1995). Om de significantie van de DIF parameters na te gaan (tweede stap) maken we gebruik maken van de WALD tests die geïmplementeerd zijn in de SAS procedure NLMIXED of van Bayesiaanse tests die steunen op de output van het Metropolis algoritme.

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 van een test voor transitief redeneren. 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.

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).

Het verklaren van DIF wil zeggen dat men verschillen in moeilijkheidsgraden of verschillen in discriminatiegraden 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 of discriminatiegraden en moeilijkheidsgraden zelf in beide groepen of discriminatiegraden 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 of de discriminatiegraad 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:

Tabel 3.1 beschrijft de modellen die het meest van belang zijn in het kader van dit project, namelijk modellen om moeilijkheidsgraden of een verschil in moeilijkheidsgraden te verklaren op basis van itemkenmerken. Model 1 gebruikt aparte parameters voor de moeilijkheidsgraden en de discriminatiegraden 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.

In principe kunnen de modellen ook uitgebreid worden om (verschillen in) discriminatiegraden te verklaren maar dit is minder van belang in het kader van het project omdat discriminatiegraden niets zeggen over de moeilijkheid van items maar wel over de samenhang van items met de onderliggende vaardigheid.

Om te evalueren of de itemkenmerken een goede beschrijving geven van de parameters die ze dienen te verklaren kan men kijken naar de significantie van de geschatte gewichten (de ηs en τs), naar de correlatie tussen de parameterwaarden van de DIF analyse en de lineaire combinatie van itemkenmerken die gebruikt worden om deze parameterwaarden te benaderen, of kan men het model statistisch vergelijken (bijvoorbeeld met een likelihood-ratio test) met het model waar geen verklaring wordt gegeven op basis van itemkenmerken. Deze laatste test is erg streng omdat de lineaire combinatie van itemkenmerken dan nagenoeg perfect de parameterwaarden van de DIF analyse dient te beschrijven.

Tabel 3.1 Modellen voor DFF

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

Er dienen nog twee kanttekeningen geplaatst te worden bij de modellen in Tabel 3.1. Ten eerste, we hebben de modellen beschreven in de veronderstelling dat er een set van ankeritems voorhanden is. Het is ook mogelijk om analoge modellen te ontwikkelen voor het geval waarbij de methode van gelijke populatiegemiddelden gebruikt wordt, maar hiervoor is momenteel geen aangepaste software voorhanden. Een eenvoudig alternatief is met lineaire regressie de geschatte DIF parameters proberen te verklaren in functie van itemkenmerken.

Ten tweede, we kiezen ervoor om bij het verklaren van DIF (modellen 2 en 3 in Tabel 3.1) de discriminatieparameters te fixeren op de waarden die eerder bekomen werden bij de DIF analyse (model 1 in Tabel 3.1). Dit doen we omdat we in eerste instantie willen weten hoe goed de itemkenmerken de moeilijkheidsgraden verklaren bij het model van de DIF analyse. Als we het model de vrijheid geven om ook nieuwe discriminatieparameters te kiezen, dan levert dit in het algemeen een beter passend model op maar de vergelijkbaarheid met de oorspronkelijke DIF analyse is zoek.

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 een 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. 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. (zie Hoofdstuk over dataverzameling). We zullen in wat volgt bespreken hoe men het effect van DIF in individuele items kan nagaan op de somscore of de gecorrigeerde 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 gemakkelijk 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}

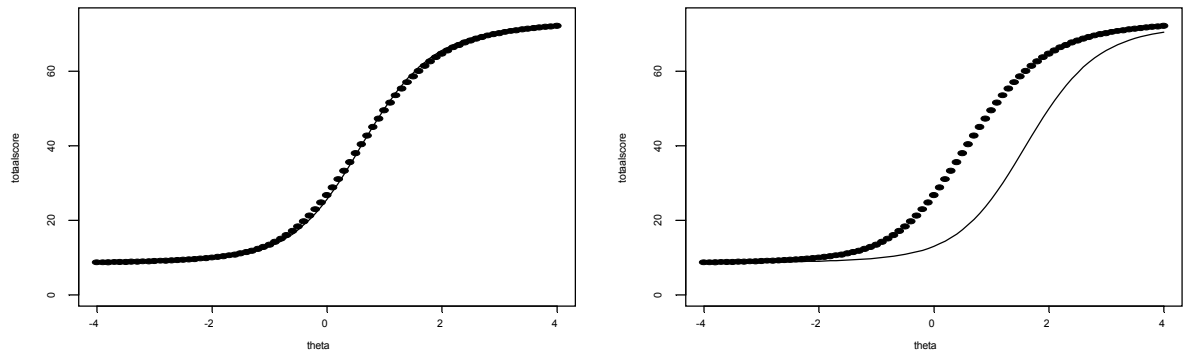
$$E(S_{pz})=\sum_i Pr(Y_{pi}=1|\theta_p,z_p)$$

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 mannen (dunne lijn) en vrouwen (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 mannen en dat een aantal andere items DIF vertonen in het voordeel van de vrouwen. 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 mannen en vrouwen omdat eenzelfde somscore kan tot stand komen door succes op verschillende verzamelingen van items die elk een tegengestelde DIF vertonen.



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

3.4.2 Gecorrigeerde somscore

De berekening van het effect van DIF in individuele items op de gecorrigeerde somscore kan niet eenvoudig analytisch benaderd worden omdat het itemresponsmodel geen onderscheid maakt tussen items die fout zijn of items die niet opgelost werden (hetzij wegens tijdsgebrek, hetzij wegens te moeilijk) en omdat het itemresponsmodel het mechanisme voor het ontbreken van gegevens niet expliciet met een kansmodel modelleert. Een eenvoudige manier om toch het effect van DIF in individuele items op de gecorrigeerde somscore te benaderen, is nagaan of het verband tussen de latente variabele en de gecorrigeerde somscore *in de geobserveerde steekproef* verschilt naargelang de groep. Dit kan concreet door elke persoon weer te geven in een grafiek met als X-as de positie van de persoon op latente variabele θ en als Y-as de gecorrigeerde somscore van de persoon. Vervolgens wordt met behulp van een "smoothing" techniek voor elke groep een vloeiende gecorrigeerde somscore-curve gefit door de punten van de personen die tot de groep behoren. Meerbepaald gebruiken we in dit rapport de "kernel smoother" met "bandwidth" parameter gelijk aan 0.5. Als de curves van beide groepen systematisch verschillen voor een bepaald deel van de latente schaal, dan heeft de DIF in individuele items in de twee groepen een differentieel effect op de gecorrigeerde somscore.

Merk op dat de methode die hier wordt voorgesteld in principe ook kan gebruikt worden om het effect van DIF op de somscore te bestuderen of bijvoorbeeld om na te gaan of de proportie personen met dezelfde positie op de schaal die een item niet invult anders is naargelang de groep.

Hoofdstuk 4: DIF-analyses

In dit Hoofdstuk analyseren we de tests van SELOR en ABL. Elke analyse bevat 3 stappen: (1) onderzoeken of individuele items voor mannen versus vrouwen DIF vertonen in de moeilijkheidsgraad of in de discriminatiegraad, (2) verklaren van DIF in functie van itemkenmerken, (3) onderzoeken wat het effect is van DIF in individuele items op de testscore en op de gecorrigeerde testscore.

Zoals uitgelegd in Hoofdstuk 3 is het bij DIF onderzoek noodzakelijk om de itemparameters die geschat werden voor elke groep (moeilijkheidsgraden en discriminatiegraden) op dezelfde schaal te plaatsen. Hiervoor zijn verschillende methoden voorhanden zoals de methode van gelijke populatiegemiddelden en de anker methode. Aangezien we op voorhand geen informatie hebben over welke items een goed anker zouden vormen wordt, gebruiken we in eerste instantie steeds de methode van gelijke populatiegemiddelden en kiezen op basis van deze resultaten een anker.

Omwille van praktische redenen bevatten de gerapporteerde DIF analyses (eerste stap) bij SELOR nog geen anker terwijl dit bij de tests van ABL al wel het geval is. Men kan er evenwel van uit gaan dat beide types van analyses (met of zonder anker) ongeveer hetzelfde resultaat opleveren omdat het anker steeds empirisch bepaald werd op basis van de resultaten van de methode met gelijke populatiegemiddelden. Bij het verklaren van DIF (tweede stap) zullen we steeds van de ankermethode gebruik maken zodat alleen DIF in niet-anker items verklaard wordt.

4.1 LOGDED

4.1.1 Modelleren van DIF

Na schatting van het 3PL op elke groep worden de itemparameters op dezelfde schaal geplaatst met de methode van gelijke populatiegemiddelden. We leggen de restrictie op dat de raadparameters dezelfde zijn in elke groep.

De verdeling van de latente variabele bij vrouwen wordt op voorhand vastgelegd ($\theta \sim N(0,1)$) terwijl de verdeling van de latente variabele voor mannen wordt geschat op basis van de data ($\theta \sim N(.27,1.49)$). We stellen vast dat mannen gemiddeld beter presteren op de test dan vrouwen ($\mu=.27$, $p<.05$) en dat de spreiding van de latente variabele groter is bij mannen.

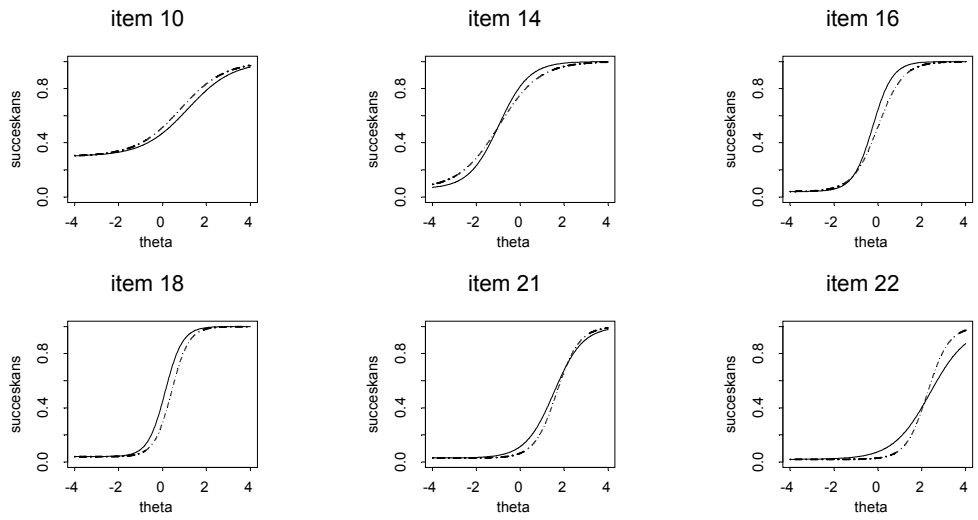
Tabel 4.1 geeft een overzicht van de geschatte itemparameters voor de vrouwen (α en β) en van de DIF die optreedt in de discriminatiegraden (ε) en in de moeilijkheidsgraden (ξ). Daarnaast geeft Tabel 4.1 per item een overzicht van de praktische significantie van de DIF aan de hand van de mediaan en het 95% BI van de absolute verschillen tussen de IRFs van mannen en vrouwen.

Tabel 4.1 Parameters van de DIF analyse. Mediaan en 95% betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor mannen en vrouwen

Item	γ	α	β	ε	ξ	Mediaan	95% BI
1	.07	.65	-3.00	.04	.37	.01	[.00,.06]
2	.08	.91	-3.34	-.14	.09	.02	[.00,.05]
3	.08	.73	-3.18	.04	.10	.00	[.00,.02]
4	.07	1.09	-2.36	-.20	-.16	.02	[.00,.07]
5	.10	.49	-1.70	-.02	-.18	.01	[.00,.02]
6	.08	1.12	-2.80	.06	.01	.00	[.00,.01]
7	.22	1.93	2.21	-.69	.00	.03	[.00,.08]
8	.07	.79	-.25	-.15	-.11	.03	[.00,.06]
9	.09	.63	-1.84	.05	-.53	.06	[.01,.08]
10	.30	.98	1.15	.02	-.33*	.03	[.00,.06]
11	.06	1.13	-2.06	-.10	.15	.02	[.00,.05]
12	.31	1.20	1.50	.18	.04	.01	[.00,.03]
13	.08	1.07	.69	-.02	.03	.00	[.00,.01]
14	.06	1.46	-.97	-.37**	.02	.04	[.00,.06]
15	.05	1.13	.66	-.01	-.10	.01	[.00,.03]
16	.04	2.23	-.23	-.48**	.24**	.01	[.00,.13]
17	.04	1.49	.84	.23	.13	.01	[.00,.07]
18	.04	2.54	.12	.00	.29**	.01	[.00,.17]
19	.03	1.80	1.50	.32	-.02	.01	[.00,.04]
20	.04	1.59	1.89	.38	-.02	.02	[.00,.05]
21	.03	1.53	1.54	.43*	.13	.02	[.00,.09]
22	.02	1.18	2.38	.86**	-.12	.04	[.00,.14]

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

Uit Tabel 4.1 blijkt dat slechts bij een beperkt aantal items DIF optreedt. Items 10 en 18 vertonen uniforme DIF respectievelijk in het voordeel van mannen ($\xi<0$) en vrouwen ($\xi>0$). Verder zien we dat items 14, 16, 21 en 22 niet-uniforme DIF vertonen. Items aan het eind van de test vertonen een sterkere samenhang met de latente trek voor mannen. Ter illustratie worden items met statistisch significante DIF ook weergegeven in Figuur 4.1. Uit de verdeling van de absolute verschillen tussen IRFs van mannen en vrouwen blijkt dat de praktische significantie van de DIF eerder beperkt is. Voor items die DIF vertonen varieert de mediaan van absolute verschillen in succeskansen van .01 tot .06 en het 97.5 percentiel varieert van .06 tot .17. 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.



Figuur 4.1 IRFs van mannen (-.-) en vrouwen (-) voor items die significante DIF vertonen ($p<.05$)

4.1.2 Verklaren van DIF

Om DIF in de moeilijkheidsgraden (ξ parameters) te verklaren kiezen we 5 anker items. Deze worden in Tabel 4.1 vet gedrukt weergegeven. De DIF in de niet-anker items wordt gemodelleerd als een lineaire functie van itemkenmerken F_1 t.e.m. F_k, namelijk $\xi_i=\Sigma\eta_k F_{ik}$ (zie ook Tabel 3.2). Op basis van een cognitieve analyse van de items werden 5 itemkenmerken onderscheiden, namelijk:

- 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 4.2 beschrijft de geschatte parameters van het model (η) en de overschrijdingskans (p) voor de toets van de nulhypothese dat de parameter gelijk is aan nul. Een p-waarde kleiner dan .05 wil bijvoorbeeld zeggen dat het gewicht van het itemkenmerk significant verschilt van nul op significantieniveau .05. De geschatte verdeling van θ voor mannen in dit model is θ~N(.34,1.20).

Tabel 4.2 Samenhang van DIF in moeilijkheidsgraden en itemkenmerken bij LOGDED

Kenmerk	Omschrijving	η	p
F_1	Aantal premissen	.07	<.0001
F_2	Aantal keer ≤ in premissen	-.03	.47
F_3	Aantal vereiste omwisselingen van premissen	-.00	.88
F_4	Aantal vereiste omwisselingen binnen premissen	-.02	.69
F_5	≤ in juist alternatief	-.13	.0026

Uit Tabel 4.2 blijkt dat twee itemkenmerken een significante maar tegengestelde samenhang vertonen met de DIF in Tabel 4.1. Items met meer premissen blijken moeilijker voor mannen dan voor vrouwen met dezelfde positie op de schaal en items waarbij ≤ voorkomt in het juiste antwoordalternatief blijken moeilijker voor vrouwen. De correlatie tussen de geobserveerde DIF in de moeilijkheidsgraden (ξ in Tabel 4.1) en de DIF zoals die voorspeld wordt op basis van de 5 itemkenmerken is gelijk aan .48 wat wil zeggen dat 23% van de variantie in de DIF parameters kan verklaard worden op basis van de itemkenmerken. Het grootste deel van de variantie in de DIF parameters blijft dus onverklaard.

In een volgende analyse onderzoeken we in welke mate de moeilijkheidsgraden van niet-anker items in elke groep kunnen verklaard worden op basis van itemkenmerken.
Tabel 4.3 toont de resultaten van het betreffende model (zie derde lijn Tabel 3.1) . De geschatte verdeling van θ voor mannen in dit model is θ~N(.39,1.02).

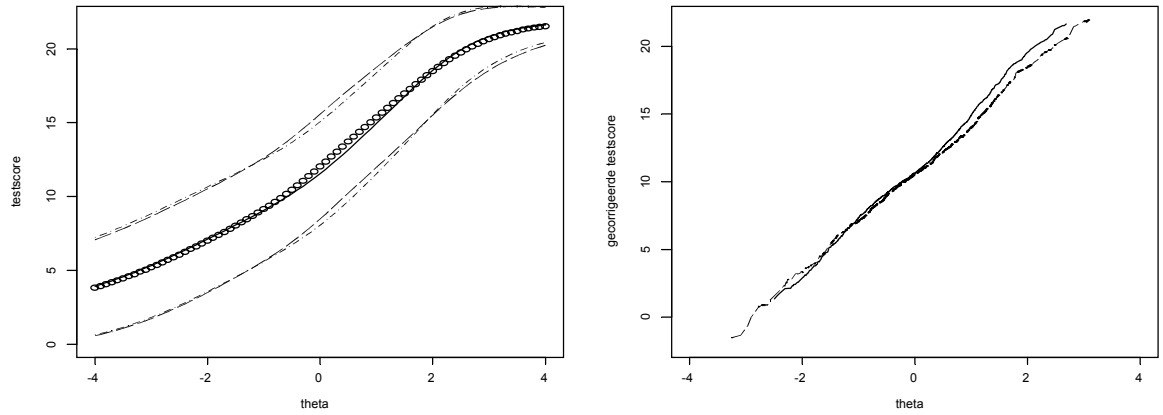
Tabel 4.3 Samenhang van moeilijkheidsgraden per groep en itemkenmerken voor LOGDED

Kenmerk	Omschrijving	τ	p	η	p
F_0	Constante	-3.3	<.0001		
F_1	Aantal premissen	.35	<.0001	.16	<.0001
F_2	Aantal keer ≤ in premissen	1.71	<.0001	-.04	.1926
F_3	Aantal vereiste omwisselingen van premissen	.44	<.0001	-.16	<.0001
F_4	Aantal vereiste omwisselingen binnen premissen	.68	<.0001	.02	.70
F_5	≤ in juist alternatief	.68	<.0001	-.29	<.0001

De positieve τ gewichten voor elk itemkenmerk geven aan dat items moeilijker worden als ze meer van deze kenmerken hebben. Bijvoorbeeld, items worden moeilijker als ze meer premissen hebben, als er meer omwisselingen nodig zijn binnen premissen enz. Daarnaast blijkt uit de η gewichten in Tabel 4.3 dat sommige itemkenmerken een verschillend gewicht hebben voor mannen en vrouwen. Items waarvoor er meer omwisselingen van premissen nodig zijn of met een $\leq$ in het juiste alternatief zijn relatief moeilijker voor vrouwen. Items met meer premissen zijn relatief moeilijker zijn voor mannen. We merken hierbij nog op dat in het model de itemgewichten gewogen worden met de discriminatieparameter zodat het effect van de itemkenmerken op de succeskans in het algemeen groter is voor items met een hogere discriminatiegraad. Uit de correlaties tussen geobserveerde en voorspelde moeilijkheidsgraden blijkt dat zowel voor mannen als voor vrouwen de moeilijkheidsgraden in elke groep redelijk goed kunnen voorspeld worden op basis van de 5 itemkenmerken. Voor mannen en vrouwen bedraagt deze correlatie telkens .88 wat wil zeggen dat 77% van de variantie in de moeilijkheidsgraden gevat wordt door het model.

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 en de gecorrigeerde testscore voor mannen en vrouwen 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 mannen en vrouwen.
De (verwachte) gecorrigeerde somscores blijken 1 punt hoger voor vrouwen met een hoge vaardigheid dan voor mannen met dezelfde hoge vaardigheid. Dit verschil is echter niet significant.



Figuur 4.2 Verwachte testscore voor mannen (-) en vrouwen (○) en 95% betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ . _) (linkerpaneel); Gecorrigeerde testscore voor mannen (_ _) en vrouwen (__) (rechterpaneel)

4.1.4 Conclusie

We stellen vast dat in de huidige steekproef mannen gemiddeld beter presteren op deze redeneertest dan vrouwen. Het geschatte vaardigheidsverschil tussen mannen en vrouwen die gemiddeld presteren bedraagt .27. Het belang hiervan voor de succeskansen van mannen en vrouwen die gemiddeld presteren kan als volgt geïllustreerd worden: Stel dat voor een bepaald item γ_i=0, β_i=0 en α_i=1.5, ε_i=0 en ξ_i=0 dan heeft de gemiddelde vrouw (θ=0) een succeskans gelijk aan 0.50 en de gemiddelde man (θ=.27) een succeskans gelijk aan .60.

Verder stellen we vast dat er voor een beperkt aantal items (6 in 22) uniforme of niet-uniforme DIF optreedt. De praktische significantie van de DIF is in het algemeen beperkt maar is voor enkele items op bepaalde stukken van de schaal wel van betekenis. De vastgestelde DIF in de moeilijkheidsgraad kan slechts in beperkte mate verklaard worden op basis van itemkenmerken (23% van de variantie in ξ parameters kan worden verklaard) maar de moeilijkheidsgraden van elke groep kunnen wel redelijk goed gevat worden in termen van een beperkt aantal itemkenmerken. (77% van de variantie in moeilijkheidsgraden van elke groep kan verklaard worden). Meerbepaald blijkt dat items met meer premissen relatief moeilijker zijn voor mannen en dat items met ≤ als juiste antwoordalternatief relatief moeilijker zijn voor vrouwen. Aangezien DIF in individuele items, afhankelijk van de samenstellende itemkenmerken, niet steeds in het voordeel is van dezelfde groep wordt ze gecompenseerd op het niveau van de gehele test en heeft ze geen differentieel effect op de (verwachte) testscores en gecorrigeerde testscores van mannen en vrouwen met dezelfde positie op de schaal.

4.2 ANAVERB

4.2.1 Modelleren van DIF

De gegevens van mannen en vrouwen werden apart geanalyseerd met het 3PL model dat gelijke raadparameters veronderstelt in elke groep. Daarna werden de parameters op één schaal geplaatst volgens de methode van gelijke populatiegemiddelden. Items 54 en 59 werden niet opgenomen in de analyse omdat de parameters niet betrouwbaar kunnen geschat worden (grote standaardfouten).

De verdeling van de latente variabele bij vrouwen werd op voorhand vastgelegd ($\theta \sim N(0,1)$). Bij de mannen werd de verdeling van de latente variabele bepaald op basis van de gegevens ($\theta \sim N(.42,1.33)$). We stellen vast dat in deze steekproef mannen gemiddeld beter presteren op de test ($\mu=.42$, $p<.01$).

Tabel 4.4 toont de geschatte itemparameters voor vrouwen (α en β) en geschatte DIF in de moeilijkheidsgraden (ξ) en in de discriminatiegraden (ε) voor mannen versus vrouwen. De ξ parameters geven dus aan hoeveel moeilijker een item is voor mannen dan voor vrouwen en de ε parameters tonen hoeveel meer de items discrimineren voor mannen dan voor vrouwen. 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 mannen en vrouwen.

We stellen vast dat ongeveer de helft van de items (50 van de 98) in de test DIF vertoont op het 5% niveau. Bij 15 items is er alleen DIF in de moeilijkheidsgraad (uniforme DIF) en bij 35 items verschilt ook de discriminatiegraad in de twee groepen (niet-uniforme DIF). De mediaan van de absolute verschillen tussen IRFs van mannen en vrouwen bij items met DIF varieert van 0 tot .11 en heeft een gemiddelde waarde van .04. Het 97.5 percentiel van de absolute verschillen varieert tussen .05 en .32 en heeft een gemiddelde waarde van .15. We kunnen besluiten dat de praktische significantie van de DIF over het algemeen beperkt is maar dat bij bepaalde items er sterke verschillen in succeskansen kunnen optreden voor een bepaald deel van de latente schaal. Ter illustratie toont Figuur 4.3 de 10 items waar het 97.5 percentiel van de absolute verschillen groter is dan .20.

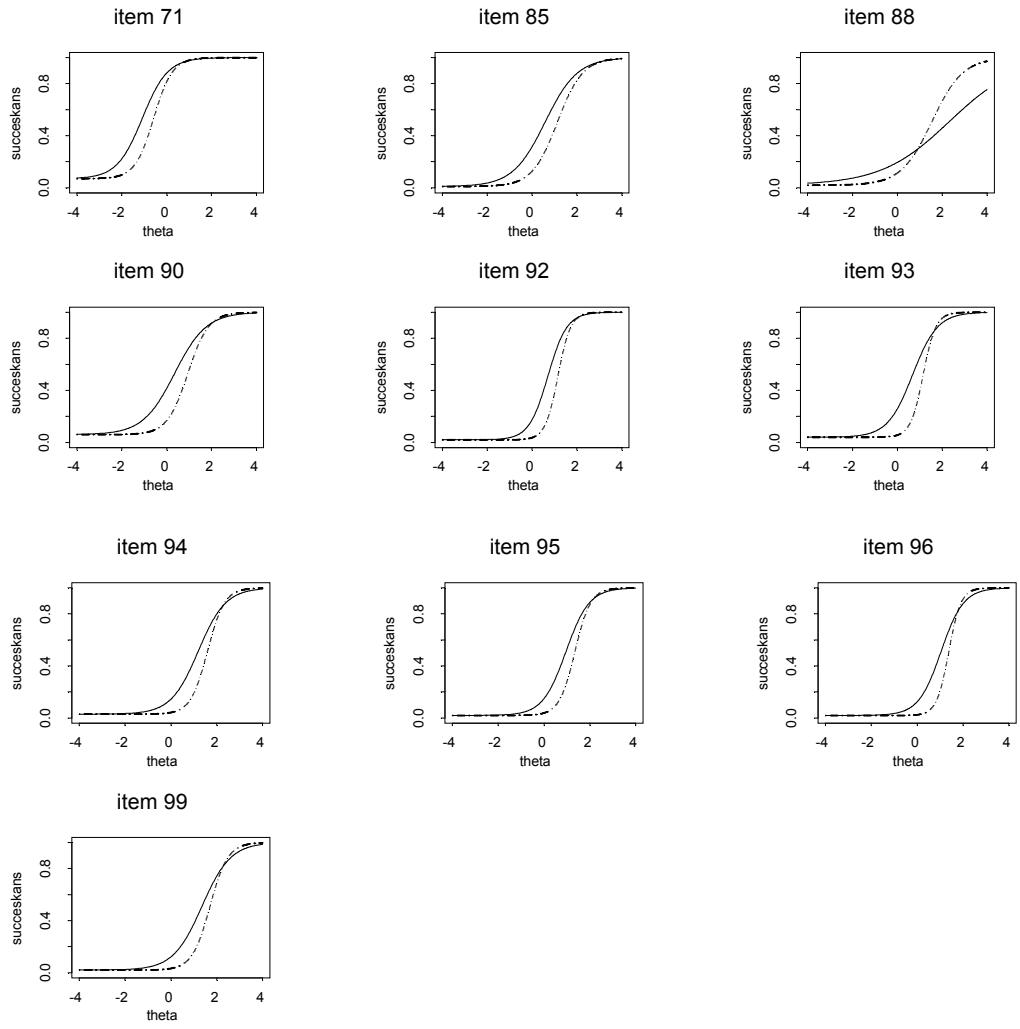
Tabel 4.4. Parameters van de DIF analyse. Mediaan en 95% betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor mannen en vrouwen

Item	γ	α	β	ε	ξ	Mediaan	95% BI
1	.03	1.06	-2.73	-.11	-.19	.01	[.00,.06]
2	.03	1.08	-3.10	-.26*	.09	.03	[.00,.07]
3	.03	.78	-2.49	-.14	-.84*	.02	[.00,.16]
4	.03	.97	-4.27	-.03	-.11	.00	[.00,.02]
5	.03	1.04	-2.46	-.28*	-.26	.03	[.00,.10]
6	.03	.46	-4.15	.01	.16	.00	[.00,.02]
7	.03	.69	-2.57	.02	.04	.00	[.00,.01]
8	.13	.93	.50	-.36**	.26*	.07	[.01,.12]
9	.03	.90	-2.35	-.09	-.53	.02	[.00,.11]
10	.03	.71	-3.44	-.25**	-.56	.04	[.01,.08]
11	.5	1.00	1.12	-.53**	.52*	.05	[.01,.11]
12	.03	.35	-3.63	.04	1.67**	.10	[.03,.15]
13	.03	.28	-2.33	-.02	-.42	.02	[.00,.03]
14	.03	.66	-2.39	-.19*	-.89*	.02	[.00,.15]
15	.03	1.04	-3.09	-.19	-.83	.00	[.00,.20]
16	.03	.74	-2.89	-.18*	-.65	.02	[.00,.12]
17	.03	.51	-2.17	-.10	-1.02*	.04	[.00,.13]
18	.03	.38	-1.47	.02	.14	.01	[.00,.02]
19	.03	.69	-2.41	-.13	-.09	.03	[.00,.05]
20	.03	.42	-3.37	.03	.18	.01	[.00,.02]
21	.02	.83	-.03	-.21*	-.97**	.11	[.00,.19]
22	.04	.80	-2.54	-.27**	-.81*	.03	[.01,.17]
23	.03	1.10	-1.56	-.25*	-.49*	.01	[.00,.14]
24	.04	.67	-1.73	-.02	-1.25**	.11	[.01,.19]
25	.04	.80	-2.75	-.15	-.25	.02	[.00,.07]
26	.02	.59	.28	-.25**	-1.07**	.11	[.01,.19]
27	.04	1.02	-2.76	-.39**	.04	.05	[.01,.10]
28	.03	.32	-2.67	.00	1.05*	.07	[.04,.08]
29	.04	.61	-1.65	-.09	.34	.05	[.00,.07]
30	.04	.83	-1.47	-.18	-.75*	.03	[.00,.15]
31	.03	.58	-2.19	-.08	-.71	.03	[.00,.10]
32	.03	.77	-3.63	-.19	-.52	.02	[.00,.08]
33	.03	.67	-2.55	.01	.09	.01	[.00,.01]
34	.03	.79	-3.13	.01	.57	.03	[.00,.11]
35	.04	.83	-1.69	-.10	-.32	.01	[.00,.07]
36	.04	.50	-.63	-.11	-.42	.03	[.00,.08]
37	.1	1.18	-.11	-.36**	-.57**	.06	[.00,.16]
38	.03	.71	-3.19	-.11	-.13	.02	[.00,.03]
39	.03	1.64	-3.42	-.08	.23	.00	[.00,.09]
40	.03	.97	-2.65	-.21	-.08	.03	[.00,.06]
41	.03	.86	-3.40	.00	.55	.03	[.00,.11]
42	.04	1.11	-1.66	-.12	-.76**	.05	[.00,.19]
43	.03	1.54	-2.98	.22	.17	.00	[.00,.08]
44	.03	.42	-1.37	-.12	-.46	.04	[.00,.09]
45	.03	.52	-2.40	.08	.08	.02	[.00,.03]
46	.03	1.07	-3.22	-.02	.55	.03	[.00,.14]
47	.03	.81	-2.50	.01	-.10	.01	[.00,.02]
48	.03	1.12	-1.56	-.12	-.29	.01	[.00,.08]
49	.03	.42	-2.44	.06	.66	.03	[.00,.08]
50	.02	.54	-1.44	-.05	-.33	.02	[.00,.05]

Vervolg Tabel 4.4

Item	γ	α	β	ε	ξ	Mediaan	95% BI
51	.03	.97	-3.30	.20	.50	.00	[.00,.14]
52	.02	.72	-1.63	.12	-.23	.03	[.00,.06]
53	.02	.76	.66	-.24**	-.14	.07	[.01,.10]
54	/						
55	.03	.42	-2.29	-.02	.44	.04	[.02,.05]
56	.03	2.12	-2.05	-.24	-.10	.00	[.00,.06]
57	.03	1.76	-2.49	-.01	.35	.01	[.00,.15]
58	.02	.92	-2.47	-.02	.27	.03	[.00,.06]
59	/						
60	.02	1.20	-1.88	-.24*	-.11	.02	[.00,.07]
61	.01	1.16	-.81	-.05	-.19	.02	[.00,.05]
62	.02	1.06	-2.01	-.01	.42*	.04	[.00,.11]
63	.02	1.35	-2.06	-.18	.25	.02	[.00,.09]
64	.02	1.29	-1.79	-.17	.34*	.03	[.00,.11]
65	.01	1.20	-.77	-.19	.29**	.03	[.00,.10]
66	.01	1.12	-.27	-.19	.11	.03	[.00,.06]
67	.01	1.00	-.15	-.01	.18*	.02	[.00,.05]
68	.01	1.20	-.65	-.12	.33**	.03	[.00,.10]
69	.02	.97	-.33	.13	.02	.02	[.00,.03]
70	.01	.85	-.37	.09	.19	.02	[.00,.05]
71	.07	1.77	-1.08	.63**	.50**	.01	[.00,.25]
72	.01	1.15	-.46	.24*	.30**	.02	[.00,.11]
73	.01	1.18	-.36	.17	.30**	.02	[.00,.10]
74	.02	.90	-.12	.19	.11	.03	[.00,.06]
75	.01	1.49	.16	.16	.03	.01	[.00,.03]
76	.01	1.13	.78	.27*	-.10	.03	[.00,.07]
77	0	1.02	.81	-.03	.22*	.02	[.00,.06]
78	.01	1.17	-.25	.50**	.47**	.03	[.01,.19]
79	0	1.47	.38	.36**	.09	.02	[.00,.07]
80	.01	1.55	2.11	.12	-.09	.00	[.00,.04]
81	.01	.91	2.66	-.09	.11	.01	[.00,.04]
82	.01	.65	.70	.33**	.40**	.07	[.01,.14]
83	.02	1.15	.63	.55**	.17*	.04	[.01,.12]
84	0	1.48	1.17	.22	.20*	.01	[.00,.09]
85	.01	1.39	.58	.39*	.55**	.03	[.00,.22]
86	.01	1.13	.73	.27*	.51**	.04	[.00,.17]
87	.01	1.92	1.01	.25	.04	.01	[.00,.04]
88	.02	.66	2.32	.80**	-.78**	.07	[.02,.28]
89	.04	1.60	.75	.98**	.26**	.02	[.00,.18]
90	.06	1.41	.35	.76**	.59**	.02	[.00,.26]
91	.01	1.17	2.96	.54	-.13	.02	[.00,.11]
92	.02	2.37	.71	1.16**	.45**	.00	[.00,.32]
93	.04	1.80	.68	1.73**	.47**	.02	[.00,.32]
94	.03	1.67	1.20	1.03**	.40**	.02	[.00,.24]
95	.02	1.97	.96	.95**	.39**	.01	[.00,.25]
96	.02	2.04	1.04	1.72**	.34**	.01	[.00,.28]
97	.01	1.96	1.39	.58*	.28**	.01	[.00,.17]
98	.02	1.55	2.20	.50	.01	.02	[.00,.06]
99	.02	1.59	1.36	.97**	.35**	.02	[.00,.22]
100	.01	1.14	2.60	.66**	-.07	.03	[.00,.11]

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



Figuur 4.3 IRFs van mannen (-.-) en vrouwen (-) voor items waar het 97.5 percentiel van de absolute verschillen groter is dan .20.

4.2.2 Verklaren van DIF

Om de DIF in de moeilijkheidsgraden te verklaren onderzoeken we in welke mate de ξ parameters kunnen verklaard worden als een functie van itemkenmerken. Bij de verbale analogieën in de ANAVERB test dient men eerst de relatie tussen twee woorden A en B te ontdekken en vervolgens dient men deze relatie toe te passen op het woord C om het woord D te vinden (zie het voorbeeld in onderstaand kader). In een cognitieve analyse van de testitems werden de 7 types van relaties onderscheiden:

Voorbeelditem:

Lood	(A)	pluim	(B)		Veder
Zwaar	(C)	?	(D)		Mooier
					Lucht
				→	Licht

- C is het resultaat van A
(bijvoorbeeld, A=regen C=nat B=sneeuw D=glad)

- C is tegengesteld aan A
(bijvoorbeeld, A=moe C=uitgerust B=lang D=kort)

- C is een onderdeel van A
(bijvoorbeeld, A=boek C=bladzijde B=servies D=bord)

- C is een intensificatie van A
(bijvoorbeeld, A=storm C=orkaan B=ruzie D=gevecht)

- C is een element dat behoort tot de verzameling A
(bijvoorbeeld, A=dier C=hond B=voornaam D=Greet)

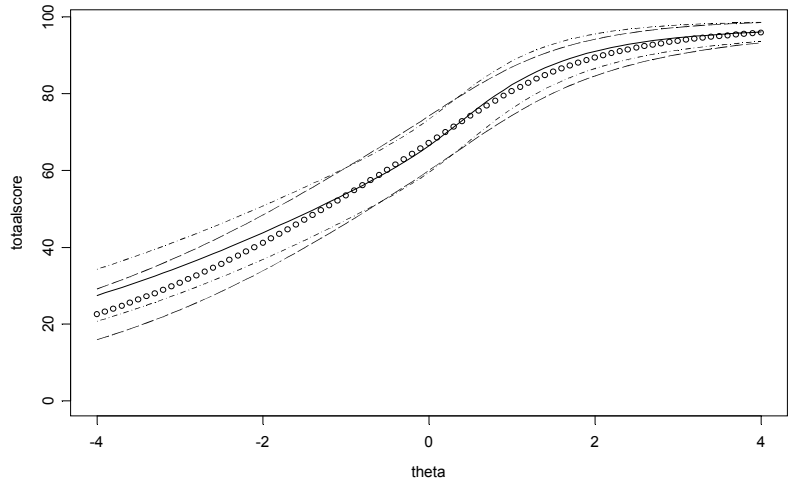
- C is een synoniem van A
(bijvoorbeeld, A=kalm C=rustig B=net D=proper)

- Tenslotte is er nog een restcategorie. Deze bevat verschillende relaties die niet tot de vorige behoren en die moeilijk te vatten zijn onder één noemer

Regressie van de ξ parameters in Tabel 4.4 op de categorische variabele die voor elk item het type van relatie aangeeft toont dat het type van relatie dat moet gezocht worden in een verbale analogie weinig samenhang vertoont met de vastgestelde DIF in de moeilijkheidsgraad. Slechts 7% van de variantie in de moeilijkheidsgraden kan worden verklaard door het type relatie. Na correctie voor het aantal predictoren in deze regressie analyse wordt zelfs slechts 1% van de variantie verklaard. Ook de moeilijkheidsgraden van de items in elke groep vertonen maar weinig samenhang met het type van relatie dat moet gezocht worden. Na correctie voor het aantal predictoren in het model verklaart het type van relatie respectievelijk 7% en 8% van de variantie in de moeilijkheidsgraden voor mannen en voor vrouwen.

4.2.3 Effect van DIF op de testscores

Om het belang van DIF op de testscores na te gaan, worden in Figuur 4.4 de verwachte somscores per groep (en bijhorend 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 mannen en vrouwen.

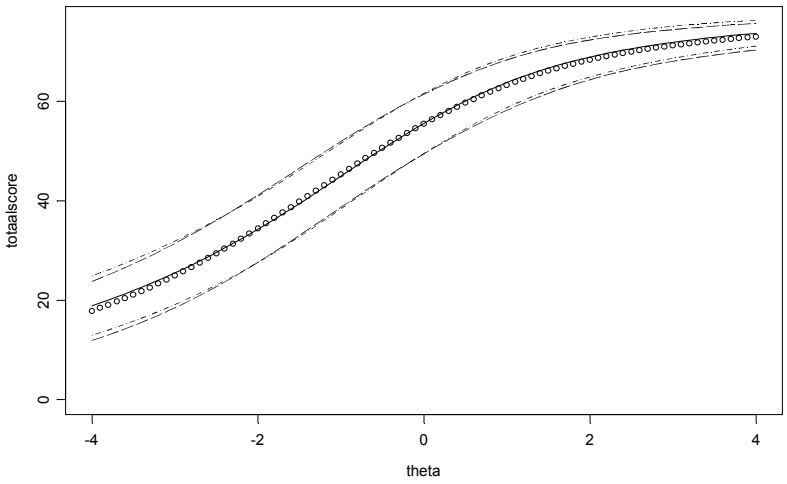


Figuur 4.4 Verwachte somscore voor mannen (-) en vrouwen (○) bij ANAVERB en 95%-betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ ._)

4.2.4 Aanpassing van ANAVERB

In tegenstelling tot wat bij andere tests werd vastgesteld, vinden we bij de ANAVERB test DIF voor een redelijk groot aantal items (50 van de 98 op het 5% niveau). Bovendien heeft de DIF in individuele items een sterker differentieel effect op de somscores van mannen en vrouwen (hoewel dit verschil strikt genomen niet significant is). Daarom lijkt het ons zinvol om na te gaan of de test kan verbeterd worden door items met sterke DIF te verwijderen. Wanneer we de 21 items, waarvan het 97.5 percentiel van de absolute verschillen groter is dan .15, uit de test verwijderen dan blijkt dat de test nog steeds een hoge betrouwbaarheid heeft (α is gelijk aan .85 en .86 voor respectievelijk mannen en vrouwen). Op deze deeltest met 76 items voeren we nogmaals de DIF analyses uit. We stellen vast dat nu nog maar 28 items DIF vertonen op het 5% niveau. Bij 14 items is er alleen DIF in de moeilijkheidsgraad (uniforme DIF) en bij 14 items verschilt ook de discriminatiegraad in de twee groepen (niet-uniforme DIF). De mediaan van de absolute verschillen tussen IRFs van mannen en vrouwen bij items met DIF varieert van .00 tot .09 en heeft een gemiddelde waarde van .02. Het 97.5 percentiel van de absolute verschillen varieert tussen .00 en .19 en heeft een gemiddelde waarde van .08. De DIF in deze kortere versie van de test is dus veel minder sterk dan in de originele test en de betrouwbaarheid van de test is nog erg hoog.

Wanneer we nu voor de deeltest het belang van DIF op de testscores onderzoeken, stellen we vast dat er geen verschil meer is tussen de verwachte somscores van mannen en vrouwen (zie Figuur 4.5). De voorgestelde aanpassing is bijgevolg succesvol geweest.



Figuur 4.5 Verwachte somscore voor mannen (-) en vrouwen (○) bij ANAVERB en 95%-betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ ._) wanneer de 21 items, waarvan het 97.5 percentiel van de absolute verschillen groter is dan .15, uit de test verwijderd worden

4.2.5 Conclusie

We stellen vast dat in deze steekproef mannen gemiddeld beter presteren op de test dan vrouwen. Verder blijkt dat ongeveer de helft van de items statistisch significante DIF vertonen. De praktische significantie van de DIF is over het algemeen beperkt (de mediaan van absolute verschillen tussen IRFs van mannen en vrouwen heeft een gemiddelde waarde van .04) maar kan voor bepaalde delen van de schaal bij bepaalde items redelijk sterk zijn. Meerbepaald geldt dat bij 10 items het 97.5 percentiel van de absolute verschillen tussen IRFs groter is dan .20 en dat bij 21 items het 97.5 percentiel van absolute verschillen tussen IRFs groter is dan .15. De DIF in de moeilijkheidsgraden vertoont weinig samenhang met het type van relatie dat dient gezocht te worden om de verbale analogie op te lossen. Tot slot blijkt dat bij lage vaardigheidsniveaus de verwachte somscores voor mannen iets hoger zijn dan voor vrouwen. Deze verschillen zijn echter niet statistisch significant zodat we kunnen besluiten dat de gezamenlijke invloed van DIF in alle items geen differentieel effect heeft op de verwachte testscores van mannen en vrouwen. Wanneer de 21 items, waarvan het 97.5 percentiel van de absolute verschillen groter is dan .15, uit de test verwijderd worden, blijkt de test nog steeds betrouwbaar te zijn, maar minder DIF te bevatten. Het verschil in verwachte somscores voor mannen versus vrouwen wordt volledig opgeheven in deze nieuwe test.

4.3 CODES

4.3.1 Modelleren van DIF

Het 3PL met gelijke raadparameters per groep wordt geschat voor mannen en vrouwen en vervolgens worden de parameters op één schaal geplaatst met de methode van gelijke populatie gemiddelden.

De verdeling van de latente variabele bij vrouwen wordt op voorhand vastgelegd $\theta \sim N(0,1)$ terwijl de verdeling van de latente variabele bij mannen bepaald wordt op basis van de gegevens ($\theta \sim N(-.09,1.09)$). Het resultaat van de analyse toont dat er geen significant hoofdeffect is ($\mu=-.09$, $p>.05$), wat wil zeggen dat mannen en vrouwen gemiddeld even goed scoren op de test.

Tabel 4.5 geeft een overzicht van de geschatte itemparameters bij vrouwen (α en β) en van de DIF in de moeilijkheidsgraden ξ en in de discriminatiegraden ε. 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 mannen en vrouwen.

Tabel 4.5 Parameters van de DIF analyse. Mediaan en 95% betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor mannen en vrouwen

Item	γ	α	β	ε	ξ	Mediaan	95% BI
1	.03	.34	-.77	.02	.04	.01	[.00,.01]
2	.03	.61	-2.00	-.12	-.58	.01	[.00,.10]
3	.03	2.25	1.58	-.23	.12	.00	[.00,.07]
4	.03	1.82	1.11	.60	-.14	.01	[.00,.11]
5	.03	2.54	.13	-.25	.06	.00	[.00,.04]
6	.03	2.89	.21	-.58	.03	.01	[.00,.06]
7	.03	2.62	-.33	-.42	.07	.01	[.00,.06]
8	.13	2.48	1.21	.11	-.13	.00	[.00,.08]
9	.03	3.08	-.02	-.50	.13	.00	[.00,.10]
10	.03	2.93	-.64	-.08	.10	.00	[.00,.07]
11	.50	1.78	.15	-.06	-.03	.00	[.00,.02]
12	.03	1.88	.69	.67*	-.05	.01	[.00,.08]
13	.03	2.39	-.23	-.01	-.06	.00	[.00,.03]
14	.03	2.82	-.31	-.53	.11	.01	[.00,.09]
15	.03	1.84	.69	.15	.08	.00	[.00,.05]
16	.03	2.02	-.33	-.25	.23*	.01	[.00,.11]
17	.03	2.57	.16	-.34	.14	.00	[.00,.09]
18	.02	2.20	.26	.24	.20	.00	[.00,.12]
19	.03	2.90	-.32	-.68*	-.02	.01	[.00,.06]
20	.03	.76	3.03	-.17	.22	.03	[.00,.06]
21	.03	2.46	1.29	-.11	-.06	.00	[.00,.04]
22	.04	.68	2.75	-.16	.80	.01	[.00,.13]
23	.04	1.40	1.52	-.05	-.12	.01	[.00,.04]
24	.03	1.71	1.26	.20	-.15	.00	[.00,.07]

$p<.05$, ** $p<.01$

Vervolg Tabel 4.5

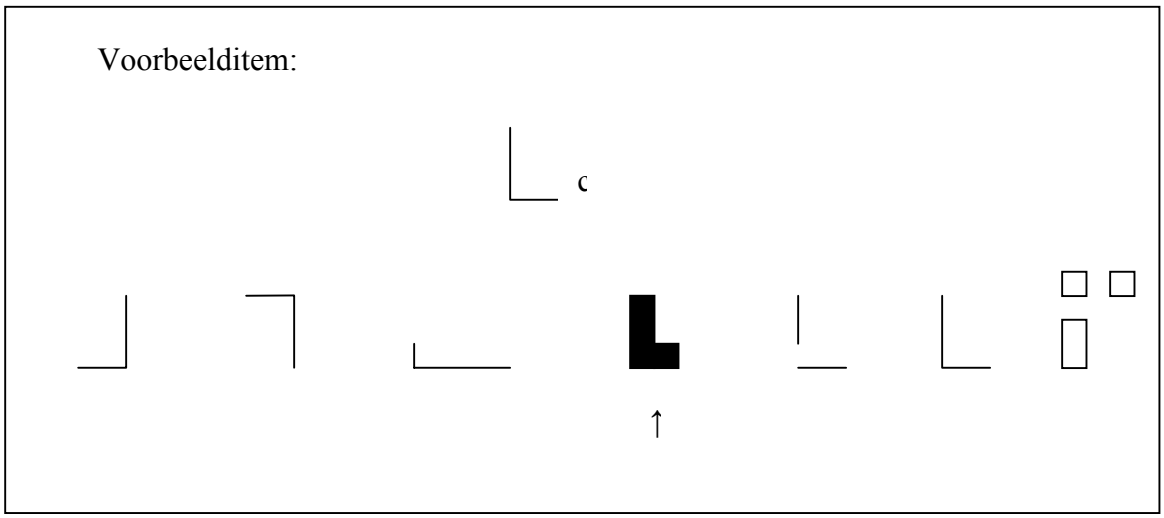
Item	γ	α	β	ε	ξ	Mediaan	95% BI
25	.04	1.84	1.30	.21	-.01	.01	[.00,.03]
26	.02	1.09	.76	-.17	.02	.02	[.00,.04]
27	.04	2.50	.33	.14	.14	.00	[.00,.09]
28	.03	.95	.90	-.02	.09	.01	[.00,.02]
29	.04	1.21	-.98	.40	.26	.02	[.00,.12]
30	.04	2.09	.72	.51	.12	.01	[.00,.09]
31	.03	3.13	.78	-.21	-.05	.00	[.00,.04]
32	.03	2.14	1.05	.22	.09	.00	[.00,.06]
33	.03	2.55	.51	.76	.12	.00	[.00,.11]
34	.03	3.32	.47	.26	-.02	.00	[.00,.03]
35	.03	2.65	.84	.33	.15	.00	[.00,.11]
36	.03	2.10	.35	-.01	.01	.00	[.00,.00]
37	.04	2.32	-.01	.81*	.03	.01	[.00,.07]
38	.04	1.60	.41	-.07	.09	.00	[.00,.04]
39	.10	.46	3.01	.01	-.49	.03	[.00,.06]
40	.03	2.11	.62	.16	.12	.00	[.00,.07]
41	.03	1.17	-.04	.04	.22	.02	[.00,.06]
42	.03	2.65	.46	-.62	.06	.01	[.00,.08]
43	.02	1.81	-.08	.15	.07	.00	[.00,.04]
44	.03	2.03	.67	-.31	.00	.01	[.00,.04]
45	.04	.88	2.75	-.19	.19	.02	[.00,.07]
46	.03	3.00	.69	-.25	.12	.00	[.00,.09]
47	.03	1.34	.00	.19	.12	.01	[.00,.06]
48	.03	1.75	2.07	.07	-.41	.02	[.00,.17]
49	.03	2.06	1.86	-.35	-.27	.01	[.00,.14]
50	.02	2.25	.11	.16	-.02	.00	[.00,.02]
51	.03	1.02	1.57	.22	-.30	.01	[.00,.10]
52	.02	1.98	.63	-.30	-.12	.01	[.00,.07]
53	.03	2.38	.56	-.06	.06	.00	[.00,.03]
54	.02	2.81	.46	.62	.07	.00	[.00,.08]
55	.01	2.08	.60	.58	-.05	.01	[.00,.07]
56	.02	2.93	.92	-.87	-.08	.01	[.00,.10]
57	.03	1.32	1.11	.51	-.26	.02	[.00,.14]
58	.01	3.87	.37	-.75	-.10	.00	[.00,.10]
59	.01	2.46	1.30	-.03	-.03	.00	[.00,.02]
60	.01	2.72	1.44	-.46	.05	.01	[.00,.06]
61	.03	2.04	1.20	.05	.10	.00	[.00,.05]
62	.00	2.13	-.13	.52	.09	.01	[.00,.08]
63	.03	1.80	.03	.10	.09	.00	[.00,.04]
64	.01	1.45	1.33	.48	-.50**	.02	[.00,.22]
65	.02	.60	2.17	.05	-.43	.03	[.00,.07]
66	.02	.93	.48	.20	-.08	.03	[.00,.05]
67	.02	1.72	.38	.16	.04	.01	[.00,.03]
68	.01	1.04	.31	.19	.16	.02	[.00,.07]
69	.01	3.37	.86	-.73	-.15	.00	[.00,.13]
70	.02	2.43	.16	.46	.04	.00	[.00,.05]
71	.02	2.41	1.49	-.72	-.29	.02	[.00,.18]
72	.01	1.98	.90	-.17	.24	.01	[.00,.11]
73	.07	1.64	1.86	.28	-.24	.00	[.00,.11]
74	.02	2.11	1.38	-.08	-.05	.00	[.00,.03]

p<.05, ** p<.01

Uit Tabel 4.5 blijkt dat er slechts bij zeer weinig items DIF optreedt (5 van de 74). De praktische significantie van de DIF is ook zeer beperkt. De mediaan van de absolute verschillen tussen IRFs is gemiddeld .01 en is nooit groter dan .03. Het 97.5 percentiel van de absolute verschillen in succeskansen is over het algemeen ook tamelijk klein. Het varieert bij alle items van .00 tot .22 en is slechts bij 3 items groter dan .15.

4.3.2 Verklaren van DIF

In de test wordt gebruik gemaakt van 18 verschillende codes. Elk item bevat één of meerdere codes. De codes moeten gedecodeerd worden om tot het juiste antwoord te komen. Wanneer een code voor de eerste keer wordt aangeboden, moet de kandidaat raden wat de code zou kunnen betekenen. Vervolgens krijgt hij feedback, zodat de betekenis van de code geleerd kan worden. Onderstaand voorbeeld toont dat code 'c' wil zeggen dat de figuur/letter moet ingekleurd worden'.



Om DIF in de moeilijkheidsgraad te verklaren onderzoeken we of de ξ parameters in Tabel 4.5 kunnen gemodelleerd worden in functie van itemkenmerken. Inspectie van de items en de DIF in de items resulteerde in 5 itemkenmerken**:**

F_1	Het aantal codes waaruit een item is opgebouwd
F_2	Er wordt een code voor de eerste keer aangeboden (F_2=1) of niet (F_2=0)
F_3	Er wordt een code aangeboden die in het vorige item voor de eerste keer werd aangeboden (F_3=1) of niet (F_3=0)
F_4	Optelsom van het aantal keren dat een code voor de tweede maal wordt aangeboden
F_5	Aantal keren dat een variant van de code '%/' (= grenst aan) in het item voorkomt

Bij het modelleren van ξ parameters in functie van itemkenmerken wordt gebruik gemaakt van de anker methode. Het anker wordt in Tabel 4.5 vet gedrukt weergegeven.
Tabel 4.6 toont de samenhang tussen de DIF in de moeilijkheidsgraden en de itemkenmerken.

Tabel 4.6 Samenhang van DIF in moeilijkheidsgraden en itemkenmerken bij CODES

Kenmerk	Omschrijving	η	p
F_1	Het aantal codes waaruit een item is opgebouwd	.01	.12
F_2	Er wordt een code voor de eerste keer aangeboden	-.01	.87
F_3	Er wordt een code aangeboden die in het vorige item voor de eerste keer werd aangeboden	.12	.02
F_4	Optelsom van het aantal keren dat een code voor de tweede maal wordt aangeboden	-.05	.16
F_5	Aantal keren dat een variant van de code '%/' in het item voorkomt	-.04	.13

Uit Tabel 4.6 blijkt dat maar één itemkenmerk een significante samenhang vertoont met de DIF in Tabel 4.5. Items waarbij een code voor de tweede keer wordt aangeboden, direct na een item waar de code voor de eerste keer werd aangeboden, blijken makkelijker voor vrouwen dan voor mannen. Er is een correlatie van .39 tussen de geobserveerde DIF in de moeilijkheidsgraden (ξ in Tabel 4.5) en de DIF zoals die voorspeld wordt op basis van de 5 itemkenmerken (η in Tabel 4.6). Dit wil zeggen dat slechts 15% van de variantie in de DIF parameters verklaard kan worden op basis van de itemkenmerken.

Tabel 4.7 Samenhang van moeilijkheidsgraden per groep en itemkenmerken voor CODES

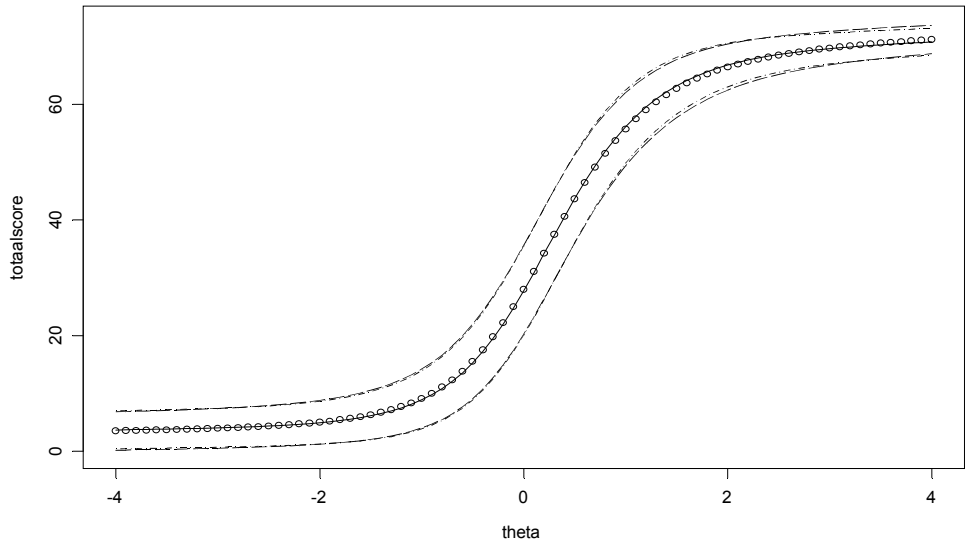
Kenmerk	Omschrijving	τ	p	η	p
F_0	constante	-.10	.001		
F_1	Het aantal codes waaruit een item is opgebouwd	.13	<.0001	.01	.20
F_2	Er wordt een code voor de eerste keer aangeboden	.92	<.0001	-.04	.18
F_3	Er wordt een code aangeboden die in het vorige item voor de eerste keer werd aangeboden	-.15	<.0001	.02	.64
F_4	Optelsom van het aantal keren dat een code voor de tweede maal wordt aangeboden	.26	<.0001	.005	.90
F_5	Aantal keren dat een variant van de code '%/' in het item voorkomt	.64	<.0001	-.23	<.0001

Tabel 4.7 toont de resultaten van de analyse waarbij de moeilijkheidsgraden (van niet-anker items) in elke groep gemodelleerd worden in functie van de 5 itemkenmerken. De geschatte verdeling van θ voor mannen is θ~N(.10;.84).
De significante positieve τ gewichten voor de itemkenmerken F_1, F_2, F_4 en F_5 geven aan dat items moeilijker worden als ze meer van deze kenmerken hebben. Bijvoorbeeld, items worden moeilijker als ze meerdere codes bevatten, als een code voor de eerste keer voorkomt, enz. Het negatieve gewicht van F3 wil zeggen dat items gemakkelijker zijn als er een code voorkomt die de eerste keer in het vorige item werd aangeboden. Daarnaast blijkt uit de η gewichten in Tabel

4.7 dat één itemkenmerk een verschillend gewicht heeft voor mannen versus vrouwen. Items waar een variant van de code '%\' in voorkomt, zijn relatief moeilijker voor vrouwen dan voor mannen. Uit de correlaties tussen geobserveerde en voorspelde moeilijkheidsgraden blijkt dat zowel voor mannen als voor vrouwen de moeilijkheidsgraden in elke groep slechts in beperkte mate kunnen voorspeld worden op basis van de 5 itemkenmerken. Voor mannen bedraagt deze correlatie 0.58 en voor vrouwen 0.59. De itemkenmerken verklaren dus respectievelijk 34% en 35% van de variantie in de moeilijkheidsgraden van de items.

4.3.3 Effect van DIF op de testscores

Om het belang van DIF op de testscores na te gaan, worden in Figuur 4.6 de verwachte somscores (en bijhorend betrouwbaarheidsinterval) per groep 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 mannen en vrouwen.



Figuur 4.6 Verwachte testscore voor mannen (-) en vrouwen (○) bij CODES en 95%-betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ ._)

4.3.4 Conclusie

We stellen vast dat mannen en vrouwen gemiddeld even goed presteren op de CODE test. Verder blijkt dat er slechts in zeer weinig items DIF optreedt (5 van de 74) en dat de praktische significantie van de DIF zeer beperkt is. De mediaan van de absolute verschillen tussen IRFs van mannen en vrouwen is gemiddeld .01 en het 97.5 percentiel van de absolute verschillen tussen IRFs is slechts in enkele gevallen groter dan .15. De DIF in de moeilijkheidsgraden kan maar beperkt verklaard worden op basis van itemkenmerken. We stellen vast dat items moeilijker zijn voor mannen als ze een code bevatten die voor de tweede maal werd aangeboden en die ook in

het voorgaande item aan bod kwam. Ook de moeilijkheidsgraden in elk van de groepen apart kan slechts beperkt verklaard worden op basis van de itemkenmerken.Uit deze analyses blijkt dat de itemkenmerken wel op een zinvolle manier samenhangen met de moeilijkheidsgraden en dat mannen beter zijn in items waar varianten van een specifieke code moeten geleerd worden. Tenslotte stellen we vast dat de gezamenlijke invloed van DIF in individuele items geen differentieel effect heeft op de verwachte testscores van mannen en vrouwen.

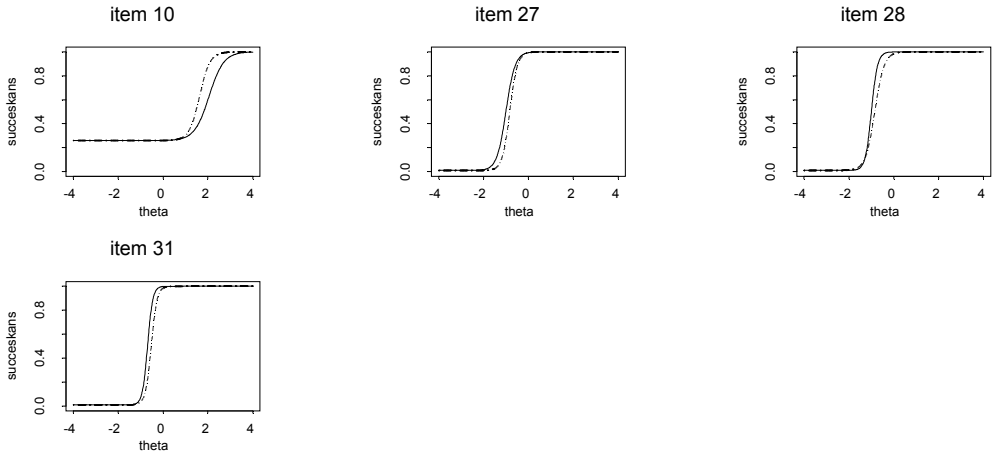
4.4 NUMVA

4.4.1 Modelleren van DIF

Na schatting van het 3PL model op elke groep, worden de parameters op één schaal geplaatst met de methode van gelijke populatiegemiddelden. De raadparameters worden verondersteld gelijk te zijn in elke groep.
De verdeling van de latente variabele bij vrouwen wordt op voorhand vastgelegd θ~N(0,1)). Bij de mannen wordt de verdeling van de latente variabele bepaald op basis van de data (θ~N (.10,1.17). We stellen vast dat in deze steekproef mannen en vrouwen gemiddeld even goed scoren op de test (μ=.10, p>.05).

Tabel 4.8 geeft een overzicht van de geschatte itemparameters bij vrouwen (α en β) en van de geschatte DIF in de moeilijkheidsgraden (ξ) en de discriminatiegraden (ε) voor mannen versus vrouwen. 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 mannen en vrouwen.

Uit tabel 4.8 kunnen we besluiten dat er bij 11 items (van de 38) DIF optreedt. Bij het merendeel (8 items) is er alleen DIF in de moeilijkheidsgraad (uniforme DIF) en bij item 20, 23 en 28 verschilt ook de discriminatiegraad in de twee groepen (niet-uniforme DIF).
De mediaan van de absolute verschillen tussen IRFs van mannen en vrouwen bij items met DIF varieert van 0 tot .08 en heeft een gemiddelde waarde van .02. Het 97.5 percentiel van de absolute verschillen varieert tussen .00 en .28 en heeft een gemiddelde waarde van .10. Absolute verschillen in succeskansen zijn in het algemeen erg klein, maar bij bepaalde items kunnen er sterke verschillen optreden in succeskansen voor een bepaald deel van de latente schaal. Ter illustratie toont figuur 4.7 de 4 items waar het 97.5 percentiel van de absolute verschillen groter is dan .20.



Figuur 4.7 IRFs van mannen (-.-) en vrouwen (-) voor items waar het 97.5 percentiel van de absolute verschillen groter is dan .20.

Tabel 4.8 Parameters van de DIF analyse. Mediaan en 95% betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor mannen en vrouwen

Item	γ	α	β	ε	ξ	Mediaan	95% BI
1	.10	.77	-4.11	-.15	-.19	.02	[.00,.03]
2	.10	.52	-3.30	-.11	-.36	.02	[.00,.05]
3	.27	3.57	2.43	-1.22	-.24	.00	[.00,.15]
4	.10	.71	-4.07	.17	.02	.02	[.00,.04]
5	.10	.54	-3.59	.10	-.37	.05	[.01,.07]
6	.10	.46	-3.69	.01	.22	.01	[.00,.02]
7	.10	.38	-2.63	.06	-.48	.05	[.03,.06]
8	.27	4.29	2.14	-1.75	-.19	.00	[.00,.16]
9	.10	.64	-4.15	.00	.02	.00	[.00,.00]
10	.26	3.07	2.05	1.28	-.42*	.00	[.00,.27]
11	.38	2.32	1.75	1.83	-.18	.01	[.00,.13]
12	.30	4.17	1.64	.46	.07	.00	[.00,.06]
13	.18	3.20	2.12	-1.42	.04	.01	[.00,.11]
14	.10	.43	-3.11	-.02	.34	.03	[.01,.03]
15	.10	.71	-3.70	.11	.00	.01	[.00,.03]
16	.10	.75	-3.09	-.06	.14	.02	[.00,.03]
17	.13	.88	2.74	.03	.26	.02	[.00,.05]
18	.09	.45	2.98	-.14	-.34	.08	[.00,.09]
19	.10	.46	-2.25	.13	.30	.03	[.00,.07]
20	.37	4.37	1.55	-2.34*	-.19	.01	[.00,.15]
21	.07	.38	-.28	.12	.69**	.07	[.00,.11]
22	.06	.88	-1.27	.10	.29	.03	[.00,.07]
23	.05	.95	-1.62	.37**	.27	.03	[.00,.12]
24	.07	.48	1.35	.09	-.32	.02	[.00,.06]
25	.04	.67	.75	.17	-.27	.03	[.00,.08]
26	.02	.73	1.82	.03	-.17	.01	[.00,.03]
27	.01	4.95	-1.00	1.05	.17*	.00	[.00,.22]
28	.01	7.40	-.99	-2.55*	.15*	.00	[.00,.23]
29	.01	.87	.78	.25	.03	.04	[.01,.06]
30	.01	1.06	2.16	.24	-.59*	.02	[.00,.18]
31	.01	8.79	-.69	-1.00	.17*	.00	[.00,.28]
32	.01	2.20	-.20	.09	.02	.00	[.00,.01]
33	.01	1.52	1.18	.22	-.22*	.01	[.00,.10]
34	.00	1.68	2.86	-.23	.46	.00	[.00,.18]
35	.01	3.04	-.01	-.37	.20**	.00	[.00,.14]
36	.01	2.50	.13	.02	.25**	.00	[.00,.15]
37	.00	2.10	.75	.04	.12	.00	[.00,.06]
38	.00	1.66	1.68	.09	.31	.02	[.00,.13]

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

4.4.2 Verklaren van DIF

Om de DIF in de moeilijkheidsgraden te modelleren maken we gebruik van lineaire regressie met als afhankelijke variabele de DIF parameters of de moeilijkheidsgraden in een bepaalde groep en als onafhankelijke variabelen bepaalde itemkenmerken. Deze aanpak is anders dan bij de andere tests omdat bij de NUMVA test geen goede set van ankeritems kon bepaald worden. De NUMVA test bestaat uit 38 cijferreeksen, die elk een aantal itemkenmerken of features bevatten. Op basis van een cognitieve analyse van de items werden volgende 6 itemkenmerken onderscheiden:

Kenmerk	Omschrijving
F_1	De cijferreeks is een doorlopende cijferreeks (F_1=1) of niet (F_1=0): De regel die moet toegepast worden, geeft een verband aan tussen twee opéénvolgende cijfers van de cijferreeks. Vb: 1 2 3 4 5 … (antw: 6) regel = +1
F_2	De cijferreeks bestaat uit twee door elkaar verweven cijferreeksen (F_2=1) of niet (F_2=0) : De even cijfers volgen één bepaalde regel (regel 1). Bij de oneven cijfers is er een andere regel van toepassing (regel 2). Vb: 1 ***3*** 2 ***6*** 3 ***12*** 4 ***24*** 5 … (antw: ***48***) regel 1 = +1, ***regel 2 = x2***
F_3	Er is geen regel in de cijferreeks maar de cijferreeks bestaat uit afzonderlijke bewerkingen (F_3=1) of niet (F_3=0) : Tussen de opéénvolgende getallen van de reeks is er geen verband. De cijferreeks bestaat uit afzonderlijke rekenopgaven. Vb: 1 3 **4** 9 7 **16** 5 6 … (antw: 11) regel = 1+3=4; 9+7=16; 5+6=?
F_4	De cijferreeks is geen rekenkundige cijferreeks maar een logische cijferreeks (F_4=1) of niet (F_4=0). De cijferreeks moet verder aangevuld worden volgens deze logica. Vb: 3 33 333 3333 33333 … (antw: 333333) regel = 3 aanvullen
F_5	Er zijn breuken aanwezig in de cijferreeks (in opgave of in antwoord) (F_5=1) of niet (F_5=0) Vb: ½ ¼ 1/8 1/16 … (antw: 1/32) regel = x ½
F_6	De regel van de cijferreeks bevat een moeilijk rekenkundige bewerking (F_6=1) of niet (F_6=0) Vb: 6 450 33750 … (antw: 2531250) regel = x 75

Tabel 4.9 toont de geschatte parameters (η) van het model, wanneer we de DIF in de moeilijkheidsgraden (ξ) trachten te verklaren met behulp van de 6 itemkenmerken. Een p-waarde kleiner dan .05 wil zeggen dat het gewicht van het itemkenmerk significant verschilt van nul op .05 niveau.

Tabel 4.9 Samenhang van DIF in moeilijkheidsgraden en itemkenmerken bij NUMVA

Kenmerk	Omschrijving	η	p
F_0	Constante	-.41	.22
F_1	Doorlopende cijferreeks	.39	.26
F_2	2 cijferreeksen in 1	.26	.41
F_3	Geen regel in cijferreeks maar afzonderlijke bewerkingen	.36	.24
F_4	Geen rekenkundige cijferreeks, maar een logische cijferreeks	.06	.66
F_5	Er zijn breuken aanwezig in de cijferreeks	.09	.41
F_6	De regel van de cijferreeks bevat een moeilijke rekenkundige bewerking	-.54	.08

Uit tabel 4.9 blijkt dat geen enkele feature een significante samenhang vertoont met de DIF in tabel 4.8. Na correctie voor het aantal itemkenmerken in het model blijkt dat het model nog maar 6% van de variantie in de DIF parameters verklaart. De DIF blijft dus grotendeels onverklaard.

In een tweede stap onderzoeken we in welke mate de moeilijkheidsgraden voor elke groep apart kunnen verklaard worden op basis van itemkenmerken. De resultaten van dit model worden weergegeven in Tabel 4.10.

Tabel 4.10 Samenhang van moeilijkheidsgraden per groep en itemkenmerken voor NUMVA

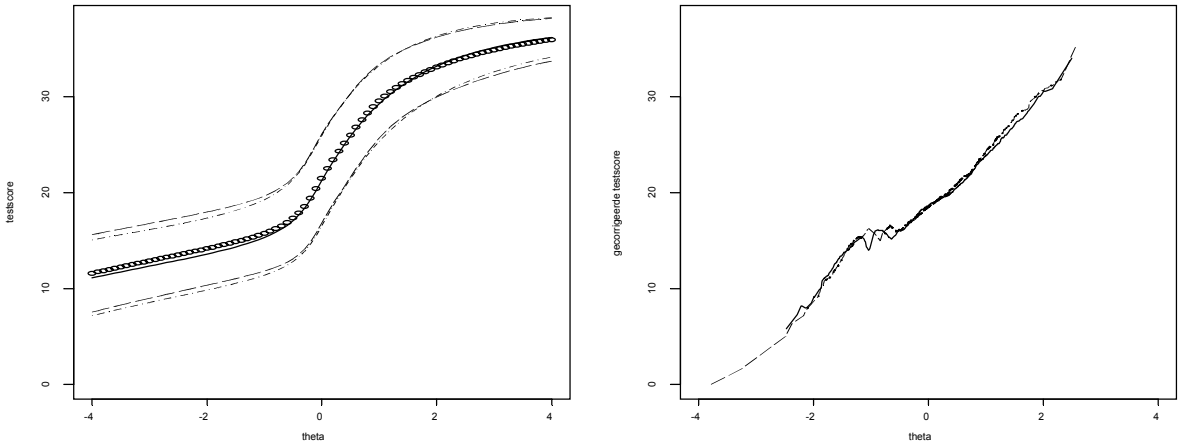
Kenmerk	Omschrijving	τ vrouwen	p	τ mannen	p
F_0	Constante	-2.40	.30	-2.81	.22
F_1	Doorlopende cijferreeks	.31	.90	.70	.77
F_2	2 cijferreeksen in 1	2.92	.20	3.18	.15
F_3	Geen regel in cijferreeks maar afzonderlijke bewerkingen	-.02	.99	.34	.87
F_4	Geen rekenkundige cijferreeks, maar een logische cijferreeks	2.26	.03	2.32	.02
F_5	Er zijn breuken aanwezig in de cijferreeks	1.18	.12	1.27	.09
F_6	De regel van de cijferreeks bevat een moeilijke rekenkundige bewerking	4.58	.04	4.04	.06

De itemkenmerk-gewichten (τ_{vrouwen} en τ_{mannen}) in Tabel 4.10 geven aan hoe de 6 itemkenmerken bijdragen aan de moeilijkheidsgraden van de items voor mannen en vrouwen. Itemkenmerken met een positief gewicht maken het item moeilijker terwijl itemkenmerken met een negatief gewicht het item gemakkelijker maken. De correlaties tussen geobserveerde en voorspelde moeilijkheidsgraden bedragen voor mannen en vrouwen respectievelijk .64 en .63, wat wil zeggen dat voor mannen 41% en voor vrouwen 40% van de variantie in de moeilijkheidsgraden wordt verklaard. Wanneer er gecontroleerd wordt voor het aantal

itemkenmerken blijkt dat voor mannen we nog maar 28% en voor vrouwen nog maar 27% van de variantie verklaren.
De τ gewichten in Tabel 4.10 hebben een zinvolle interpretatie. Zo blijken alle significante itemkenmerken het item moeilijker te maken. Wanneer de regel in een cijferreeks een logische regel is i.p.v een rekenkundige regel of wanneer er een moeilijke rekenkundige bewerking in de regel van de cijferreeks vervat zit, wordt de cijferreeks moeilijker.

4.4.3 Effect van DIF op de testscores

Figuur 4.8 toont dat er voor geen enkele waarde van θ een significant verschil is tussen de verwachte en de gecorrigeerde somscores van mannen en vrouwen. De DIF in individuele items heeft bijgevolg geen invloed op de testscores



Figuur 4.8 Verwachte testscore voor mannen (-) en vrouwen (○) en 95%-betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ ._) (linkerpaneel); Verwachte gecorrigeerde testscore voor mannen (_ _) en vrouwen (__) (rechterpaneel)

4.4.4 Conclusie

We stellen vast dat mannen en vrouwen in deze steekproef gemiddeld even goed presteren op de test. Verder blijkt dat 11 items statistisch significante DIF vertonen. 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. Voor 7 van de 38 items geldt dat het 97.5 percentiel van de absolute verschillen tussen IRFs groter is dan .15

De verschillen in de moeilijkheidsgraden kunnen niet goed verklaard worden in functie van itemkenmerken. De moeilijkheidsgraden in elke groep kunnen wel in beperkte mate en op een zinvolle manier voorspeld worden op basis van itemkenmerken. Wanneer de cijferreeks bestaat uit een logische regel of wanneer een moeilijke rekenkundige bewerking vereist is, dan wordt de cijferreeks moeilijker. Ten slotte blijkt dat de DIF in individuele items geen effect heeft op de (verwachte) somscores en gecorrigeerde somscores van mannen en vrouwen

4.5 WIMA

4.5.1 Modelleren van DIF

Uitgaande van de resultaten van de methode met gelijke populatiegemiddelden worden 10 ankeritems gekozen. Deze worden vet gedrukt weergegeven in Tabel 4.11. Vervolgens wordt in SAS het 3PL-model geschat met niet-uniforme DIF voor alle niet-anker items (zie formules (3.2) en (3.3)). Omdat volgens deze analyse er nagenoeg geen items zijn met niet-uniforme DIF, wordt een model met alleen uniforme DIF geschat met SAS. Een vergelijking van de fit van beide modellen (met niet-uniforme en uniforme DIF) aan de hand van een LR toets toont dat de nulhypothese van gelijke discriminatieparameters voor mannen en vrouwen niet kan verworpen worden (LR=21, df=13, p=.07). Tabel 4.11 beschrijft de resultaten voor het 3PL model met uniforme DIF voor niet-anker items.

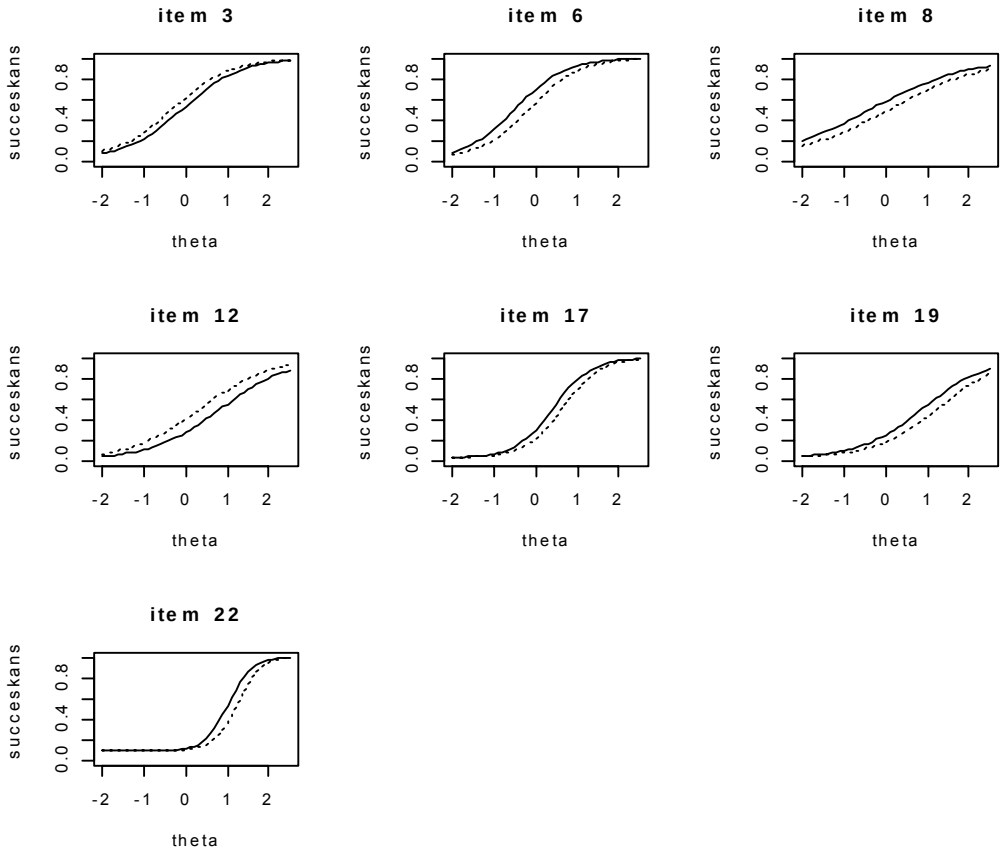
De verdeling van de latente variabele θ voor vrouwen wordt op voorhand vastgelegd in de analyse ($\theta \sim N(0,1)$). Voor mannen wordt het gemiddelde en de spreiding van de θ uit de gegevens geschat. We stellen vast dat mannen gemiddeld beter scoren dan vrouwen op de test (μ=.27, p<.0001) en dat de spreiding van de latente variabele in de twee groepen ongeveer dezelfde is (σ^2=1.05).

Tabel 4.11. Parameters van de DIF analyse. Mediaan en 95% betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor mannen en vrouwen

Item	γ	α	β	ξ	Mediaan	95% BI
1	.11	1.27	.16	-.18	.03	[.01,.05]
2	.03	.83	-.09	0	/	/
3	.03	1.46	-.06	-.24*	.05	[.01,.08]
4	.02	1.37	-.40	0	/	/
5	.01	1.08	.15	-.22	.04	[.02,.06]
6	.03	1.70	-.48	.34**	.06	[.01,.14]
7	.01	1.18	-.37	.17	.03	[.01,.05]
8	.03	.88	-.33	.47**	.08	[.04,.10]
9	.01	1.78	-.01	0	/	/
10	.01	1.19	.09	-.20	.04	[.01,.06]
11	.01	1.43	.06	0	/	/
12	.02	1.20	.84	-.48**	.09	[.03,.14]
13	.03	1.12	.55	0	/	/
14	.01	1.70	.19	-.14	.03	[.00,.06]
15	.03	1.35	.29	0	/	/
16	.01	2.13	.73	-.16	.02	[.00,.08]
17	.04	2.23	.43	.22**	.03	[.00,.12]
18	.04	1.79	.73	0	/	/
19	.04	1.34	.92	.36**	.07	[.01,.11]
20	.10	2.69	.86	0	/	/
21	.04	2.60	1.00	0	/	/
22	.10	3.57	1.02	.21*	.01	[.00,.16]
23	.04	2.40	1.44	0	/	/

* p<.05; ** p<.01; ankeritems worden vet weergegeven

Uit Tabel 4.11 blijkt dat items 3 en 12 significant gemakkelijker zijn voor mannen en dat items 6, 8, 17, 19 en 22 significant gemakkelijker zijn voor vrouwen. Uit de verdeling van de absolute verschillen tussen IRFs van mannen en vrouwen in Tabel 4.11 blijkt verder dat de praktische significantie van de DIF voor de WIMA test eerder beperkt is: De mediaan van de absolute verschillen varieert van .01 tot .09. De IRFs van items met significante DIF voor mannen en vrouwen worden ter illustratie weergegeven in Figuur 4.9.



Figuur 4.9 IRFs van mannen (- -) en vrouwen (_) voor items die significante DIF vertonen

4.5.2 Verklaren van DIF

Om DIF te verklaren onderzoeken we of verschillen in moeilijkheidsgraden bij niet-anker items (ξ in Tabel 4.11) kunnen gemodelleerd worden in functie van itemkenmerken. Een cognitieve analyse van de vraagstukken in de test leverde 8 itemkenmerken op (zie onderstaand kader). Tabel 4.12 bevat de resultaten van het model waarbij de ξ_i gemodelleerd werden als een lineaire combinatie van de 8 itemkenmerken. De geschatte verdeling van θ is ongeveer dezelfde als in de vorige analyse (namelijk, $\theta \sim N(.27,1.05)$).

Vraagstukken

8 itemkenmerken:

- F_1 = aantal evenredigheden in item. Het oplossen van een evenredigheid wil zeggen dat men op basis van de evenredigheid A/B=C/D één van de factoren (A, B, C of D) berekent als de andere drie factoren gegeven zijn. Een alternatieve voor het oplossen van de evenredigheid is via de welbekende "regel van drie". Bijvoorbeeld,…
- F_2 = aantal vergelijkingen in item. Het oplossen van een vergelijking veronderstelt dat men op basis van de relatie tussen een onbekende factor en verschillende bekende factoren de onbekende factor berekent. Bijvoorbeeld,…
- F_3 = speciale kennis vereist (F_3=1) of niet (F_3=0)
- F_4 = item bevat tijdsberekening (F_4=1) of niet (F_4=0)
- F_5 = item bevat percentageberekening (F_5=1) of niet (F_5=0)
- F_6 = oplossing van item is geen geheel getal (F_6=1) of wel (F_6=0)
- F_7 = item bevat overbodig gegeven (F_7=1) of niet (F_7=0)
- F_8 = aantal stappen nodig om item op te lossen

Tabel 4.12 Samenhang van DIF in moeilijkheidsgraden en itemkenmerken bij WIMA

Kenmerk	Omschrijving	η	p
F_1	Aantal evenredigheden	-.27	.11
F_2	Aantal vergelijkingen	-.09	.41
F_3	Speciale kennis vereist	-.06	.34
F_4	Tijdsberekening	-.33	.02
F_5	Percentageberekening	-.68	.001
F_6	Oplossing is geen geheel getal	-.43	.11
F_7	Overbodig gegeven	-.20	.28
F_8	Aantal vereiste stappen	.26	.007

Uit Tabel 4.12 blijkt dat (na controle voor andere itemkenmerken in het model) 3 itemkenmerken significant samenhangen met geschatte verschillen in moeilijkheidsgraden (namelijk $p<.05$). Items met tijdsberekening of percentageberekening blijken moeilijker voor vrouwen terwijl items die meer stappen vereisen om tot de oplossing te komen moeilijker zijn voor mannen. Dit kan als volgt geïllustreerd worden. Stel dat men bij item i één vergelijking moet oplossen, dat het item speciale kennis vereist en dat er 4 stappen nodig zijn om het item op te lossen. Als bijvoorbeeld $\gamma_i=0$, $\beta_i=0$, en $\alpha_i=1.5$ dan is de succeskans voor een vrouw met vaardigheid θ gelijk aan:

$$\exp[\alpha_i(\theta-\beta_i)]/(1+\exp[\alpha_i(\theta-\beta_i)])= \exp[1.5\theta]/(1+\exp[1.5\theta]).$$

Voor een man met vaardigheid θ is de succeskans gelijk aan:

$\exp[\alpha_i(\theta-(\beta_i-.09-.06+4(.26)))]/(1+\exp[\alpha_i(\theta-(\beta_i-.09-.06+4(.26)))])$
$=\exp[1.5(\theta-.89)]/(1+\exp[1.5(\theta-.89)])$

Bijvoorbeeld, voor een vrouw en een man met θ=0 is de voorspelde succeskans respectievelijk gelijk aan .50 en .21.
De voorgestelde itemkenmerken in Tabel 4.12 kunnen de verschillen in de moeilijkheidsgraden tamelijk goed verklaren: De correlatie tussen de ξi in Tabel 4.11 en de DIF zoals voorspeld op basis van de 8 itemkenmerken bedraagt 0.80. Dit wil zeggen dat 64% van de variantie in de 13 ξ parameters verklaard wordt door de 8 itemkenmerken.

In een volgende stap onderzoeken we in welke mate de moeilijkheidsgraden (van niet-anker items) voor elke groep apart kunnen verklaard worden op basis van itemkenmerken. De resultaten van dit model worden weergegeven in Tabel 4.13. De verdeling van de latente variabele voor mannen is in deze analyse ongeveer dezelfde als in de vorige twee analyses (namelijk, θ~N(.29,.96).

Tabel 4.13 Samenhang van moeilijkheidsgraden per groep en itemkenmerken voor WIMA

Kenmerk	Omschrijving	τ	p	η	p
F_0	Constante	.87	<.0001		
F_1	Aantal evenredigheden	.49	.0014	-.22	.19
F_2	Aantal vergelijkingen	-.77	<.0001	-.06	.59
F_3	Speciale kennis vereist	.44	<.0001	-.04	.53
F_4	Tijdsberekening	-1.23	<.0001	-.32	.03
F_5	Percentageberekening	-.40	.04	-.60	.003
F_6	Oplossing is geen geheel getal	-2.09	<.0001	-.42	.11
F_7	Overbodig gegeven	-1.33	<.0001	-.22	.24
F_8	Aantal vereiste stappen	.12	.17	.22	.02

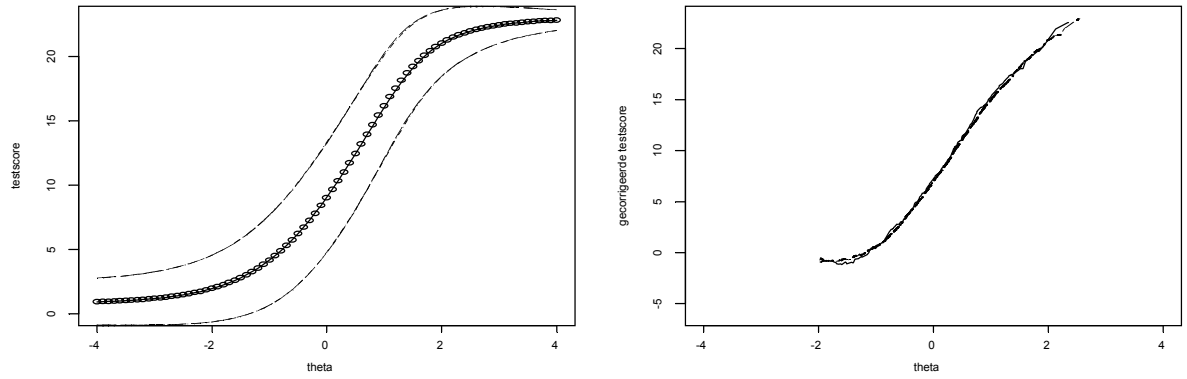
De itemkenmerk-gewichten τ in Tabel 4.13 geven aan hoe de 8 itemkenmerken bijdragen aan de moeilijkheidsgraden van de items voor vrouwen. Itemkenmerken met een positief gewicht maken het item moeilijker terwijl itemkenmerken met een negatief gewicht het item gemakkelijker maken. Daarnaast geven de η gewichten aan hoe de verschillen in de moeilijkheidsgraden kunnen verklaard worden op basis van de itemkenmerken. De correlaties tussen geobserveerde en voorspelde moeilijkheidsgraden bedragen voor mannen en vrouwen .68 en .58, respectievelijk, wat wil zeggen dat voor mannen 46% en voor vrouwen 34% van de variantie in de moeilijkheidsgraden wordt verklaard. De itemkenmerken laten dus slecht in beperkte mate toe om de moeilijkheidsgraden te voorspellen.

De τ gewichten in Tabel 4.13 hebben bovendien niet altijd een psychologisch zinvolle interpretatie. Bijvoorbeeld items waarbij men meer vergelijkingen moet oplossen zouden gemakkelijker zijn. Een mogelijke verklaring hiervoor is dat het aantal itemkenmerken relatief groot is in vergelijking met het aantal items waardoor de itemkenmerkgewichten gedeeltelijk de

ruis in de moeilijkheidsgraden weergeven. Dit probleem is moeilijk te vermijden als er relatief veel verschillende itemkenmerken een rol kunnen spelen in een beperkt aantal items.

4.5.3 Effect van DIF op de testscores

Om het belang van DIF op de testscores na te gaan, worden in Figuur 4.10 de (verwachte) somscores (en bijbehorend betrouwbaarheidsinterval) en gecorrigeerde somscores per groep in functie van θ weergegeven. De somscore-curves en gecorrigeerde 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 mannen en vrouwen. De DIF in individuele testitems heeft dus geen differentieel effect op de (verwachte) somscores en gecorrigeerde somscores van mannen en vrouwen.



Figuur 4.10 Verwachte testscore voor voor mannen (-) en vrouwen (o) en 95% betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ ._) (linkerpaneel); Verwachte gecorrigeerde testscore voor mannen (_ _) en vrouwen (__) (rechterpaneel)

4.5.4 Conclusie

We stellen vast dat mannen gemiddeld significant beter presteren op de test. Verder kunnen we stellen dat bij WIMA in beperkte mate uniforme DIF optreedt voor een relatief klein aantal items (7 van de 23 items). De verschillen in de moeilijkheidsgraden kunnen relatief goed gemodelleerd worden in functie van itemkenmerken. Deze analyses tonen dat items met tijdsberekening en percentageberekening iets gemakkelijker zijn voor mannen en dat items met een lang oplossingsproces (veel stappen) iets gemakkelijker zijn voor vrouwen. We moeten hierbij opmerken dat deze bevindingen een voorlopige waarde hebben omdat er relatief veel kenmerken gebruikt werden om de DIF in een beperkt aantal items te modelleren. Om de resultaten te valideren zou men moeten nagaan hoe goed de geschatte itemgewichten van de huidige studie de DIF voorspellen in nieuwe items die (combinaties van) dezelfde 8 itemkenmerken bevatten. De

voorspelling van de moeilijkheidsgraden van elke groep in functie van itemkenmerken was eerder matig en niet altijd psychologisch zinvol. Ook hier geldt dat de bevindingen waarschijnlijk slechts beperkt kunnen gegeneraliseerd worden naar nieuwe tests. Tot slot stellen we vast dat de DIF in individuele items geen differentieel effect heeft op (verwachte) somscores en gecorrigeerde somscores van mannen en vrouwen.

4.6 TNV

4.6.1 Modelleren van DIF

Op basis van de methode met gelijke populatie gemiddelden werden 12 ankeritems gekozen. Deze worden vet gedrukt weergegeven in Tabel 4.14. Vervolgens wordt het 3PL-model geschat met niet-uniforme DIF voor alle niet-anker items (zie formules (2.2)-(2.3)). Omdat volgens deze analyse er nagenoeg geen items zijn met niet-uniforme DIF wordt een model geschat met alleen uniforme DIF voor alle niet-anker items. Een vergelijking van de fit van beide modellen (met uniforme en niet-uniforme DIF) met een likelihood-ratio toets heeft als resultaat dat de nulhypothese van gelijke discriminatie parameters in de twee groepen niet kan verworpen worden (LR=51 ,df=38 ,p=.31).

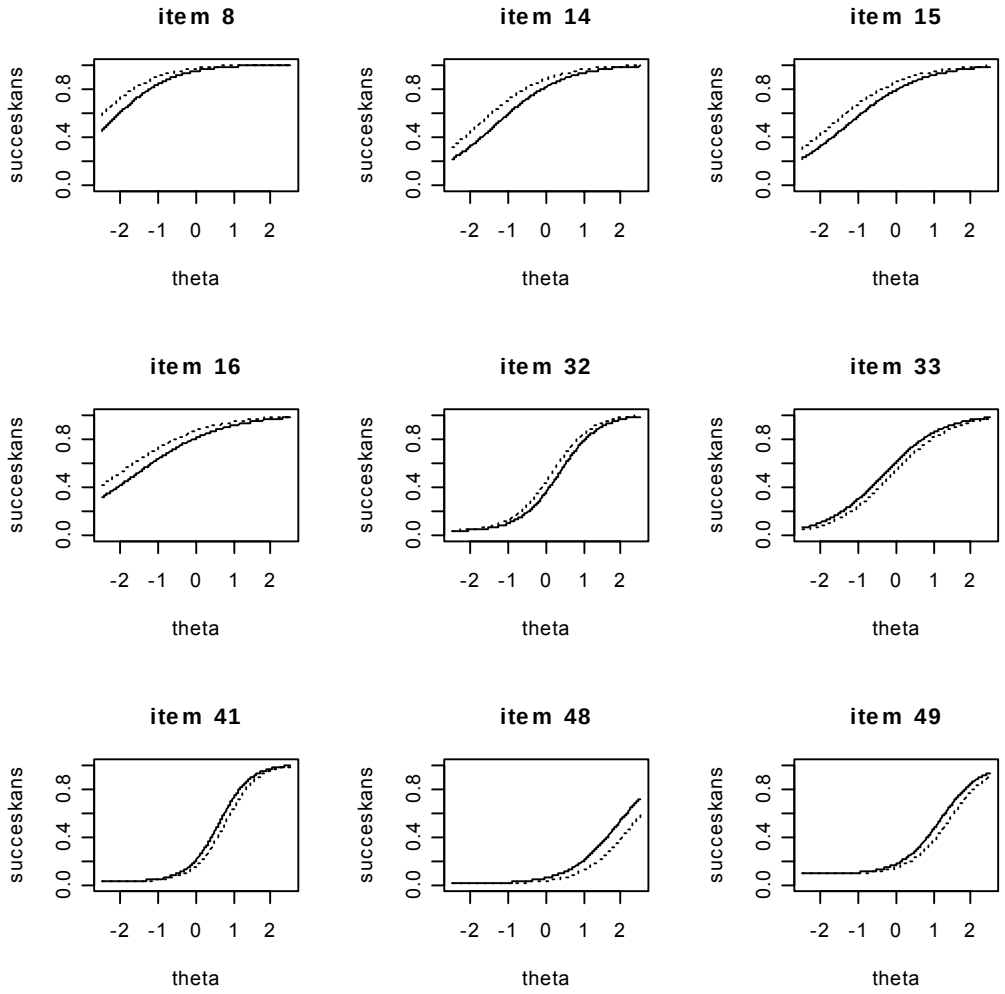
Voor vrouwen wordt de verdeling van de latente variabele vastgelegd ($\theta \sim N(0,1)$).Voor mannen wordt het gemiddelde van θ uit de gegevens geschat terwijl de variantie van θ gelijk wordt gesteld aan 1. Uit de schatting van het model met uniforme DIF blijkt dat mannen en vrouwen gemiddeld niet significant verschillend presteren op de test (μ=.08, p>.05).

Tabel 4.14 beschrijft de resultaten voor het 3PL-model met uniforme DIF voor niet-anker items. De Tabel bevat ook voor elk item de mediaan en een 95% betrouwbaarheidsinterval van de verdeling van de absolute verschillen tussen de IRFs van mannen en vrouwen.
Tabel 4.14 toont dat slechts 9 van de 38 niet-anker items uniforme DIF vertonen. Een positieve ξ wil zeggen dat het item moeilijker is voor mannen terwijl een negatieve ξ wil zeggen dat het moeilijker is voor vrouwen. Uit de verdeling van de absolute verschillen tussen IRFs van mannen en vrouwen (laatste twee kolommen van Tabel 4.14) blijkt dat de praktische significantie van de DIF zeer beperkt is. Het 97.5 percentiel van items met statistisch significante DIF schommelt meestal tussen .10 en .15, wat wil zeggen dat voor items die DIF vertonen het verschil in de succeskansen voor mannen en vrouwen zelden meer dan .15 bedraagt. Ter illustratie toont Figuur 4.11 de IRFs van mannen en vrouwen voor items met significante DIF.

Tabel 4.14 Parameters van de DIF analyse. Mediaan en 95% betrouwbaarheidsinterval (BI) van de verdeling van absolute verschillen tussen IRFs voor mannen en vrouwen

item	γ	α	β	ξ	Mediaan	95% BI
1	.03	.98	-4.96	.37	.00	[.00,.03]
2	.11	.86	-1.28	.13	.02	[.00,.02]
3	.18	1.38	-1.77	.06	.01	[.00,.02]
4	.02	.81	-3.54	.00	/	/
5	.02	1.04	-2.92	-.15	.01	[.00,.03]
6	.02	1.08	-2.29	-.21	.01	[.00,.06]
7	.02	1.25	-2.11	-.03	.00	[.00,.01]
8	.02	1.23	-2.33	-.43*	.02	[.00,.13]
9	.02	1.56	-1.48	.00	/	/
10	.01	1.03	-2.19	.08	.01	[.00,.02]
11	.01	1.28	-2.53	.17	.01	[.00,.05]
12	.01	1.29	-2.36	-.25	.01	[.00,.08]
13	.01	.92	-2.55	.00	/	/
14	.01	1.11	-1.33	-.46**	.06	[.01,.13]
15	.01	1.03	-1.28	-.42**	.06	[.01,.11]
16	.01	.88	-1.61	-.48**	.06	[.01,.10]
17	.01	1.26	-2.01	-.29	.02	[.00,.09]
18	.01	.64	-.29	-.10	.01	[.01,.02]
19	.01	.94	-1.24	.12	.02	[.00,.03]
20	.01	1.03	-1.65	.00	/	/
21	.01	.91	-1.13	-.25	.04	[.01,.06]
22	.01	.99	-1.04	.00	/	/
23	.01	.98	-.68	-.17	.03	[.01,.04]
24	.01	1.38	-.72	-.05	.01	[.00,.02]
25	.01	.61	-.55	.00	/	/
26	.01	1.29	-.96	.08	.01	[.00,.03]
27	.01	.99	-.98	.00	/	/
28	.01	.96	-.63	.14	.02	[.01,.03]
29	.01	.97	.16	.10	.02	[.01,.02]
30	.01	1.12	-.07	-.16	.03	[.01,.04]
31	.04	1.37	-.22	.03	.01	[.00,.01]
32	.04	1.91	.36	-.20**	.03	[.00,.09]
33	.01	1.29	-.34	.24*	.04	[.01,.08]
34	.00	1.60	.25	.00	/	/
35	.06	1.68	.14	.00	/	/
36	.01	1.40	-.32	.00	/	/
37	.03	1.41	.02	.05	.01	[.00,.02]
38	.01	1.58	.06	.12	.02	[.00,.05]
39	.01	1.09	-.01	-.07	.01	[.00,.02]
40	.00	1.23	.93	.00	/	/
41	.04	2.41	.62	.17*	.02	[.00,.10]
42	.01	1.73	1.21	-.09	.01	[.00,.04]
43	.07	2.28	1.30	-.17	.02	[.00,.09]
44	.06	1.75	.96	.12	.02	[.00,.05]
45	.02	2.32	1.11	.09	.01	[.00,.05]
46	.05	2.32	1.44	-.06	.00	[.00,.03]
47	.06	1.66	2.97	.00	/	/
48	.02	1.55	1.91	.41**	.02	[.00,.15]
49	.10	1.93	1.19	.24*	.03	[.00,.10]
50	.05	1.83	1.74	.17	.01	[.00,.07]

*p<.05; **p<.01; ankeritems worden vet weergegeven

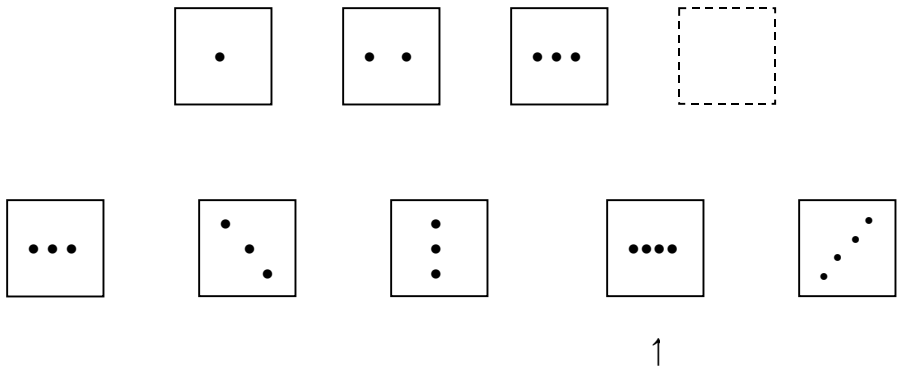


Figuur 4.11 IRFs van mannen (- -) en vrouwen (_) voor items die significante DIF vertonen

4.6.2 Verklaren van DIF

Om DIF te verklaren gaan we na in welke mate verschillen in moeilijkheidsgraden bij niet-anker items (ξ in Tabel 4.10) kunnen gemodelleerd worden in functie van itemkenmerken. Onderstaande kader toont een typisch item van de TNV test.
Vervolgens onderscheiden we de itemkenmerken die in de test aanwezig zijn en illustreren ze met een fictief voorbeeld.

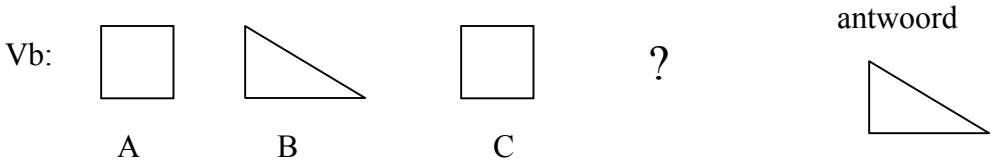
Voorbeelditem:



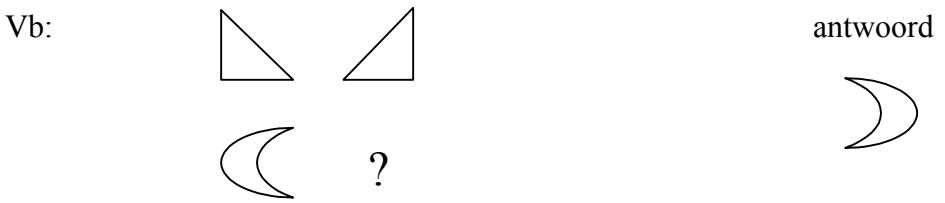
- Het aantal reeksen waaruit het item is opgebouwd (F_1):
    a) Items waar één reeks van 3 geometrische figuren aangevuld moet worden met een vierde figuur. (cf. supra: voorbeelditem)
    b) Items waar in een eerste reeks een verband tussen 2 geometrische figuren moet gezocht worden. Deze kennis moet vervolgens toegepast worden om een tweede reeks aan te vullen. (cf. infra: vb F3)
    c) Items waar eenzelfde verband in 2 reeksen van 3 geometrische figuren moet gezocht worden om een derde reeks aan te kunnen vullen. (cf. infra: vb F4)

  De drie types van items worden gecodeerd met twee binaire variabelen R_1 (gelijk aan 1 als het item bestaat uit één reeks figuren en 0 anders) en R_2 (gelijk aan 1 als het item bestaat uit twee reeksen figuren en 0 anders).

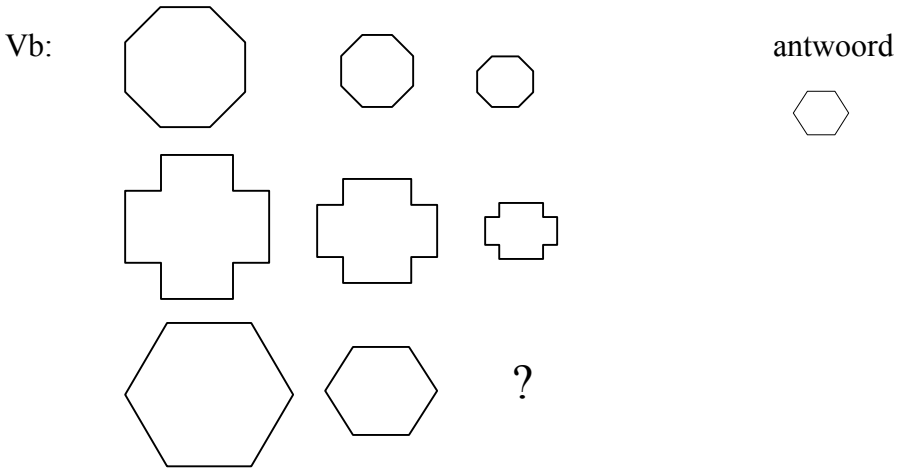
- Gelijkheid (F_2=1) of niet (F_2=0) : De geometrische figuur C is gelijk aan figuur A, waaruit volgt dat het antwoord gelijk moet zijn aan figuur B.



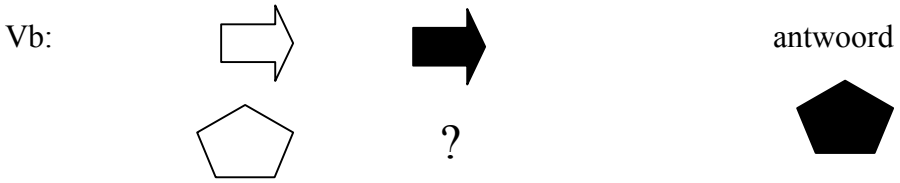
- Aantal spiegelingen (F_3): De geometrische figuur moet gespiegeld worden om tot de juist oplossing te komen.



- Aantal keer verkleinen/vergroten/verbreden/versmallen (F_4): Men moet de figuur verkleinen, vergroten, verbreden of versmallen om tot het juiste antwoord te komen.

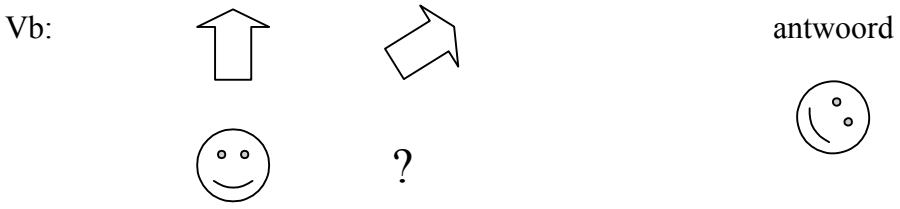


- Motiefverandering (F_5=1) of niet (F_5=0): Dit is van toepassing als de geometrische figuur een ander opvulling krijgt.

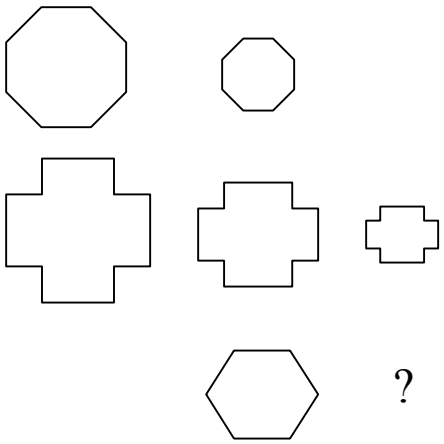


- Aantal eenvoudige bewerkingen (F_6) (cf. voorbeelditem): In de reeks van figuren bevindt zich een patroon waarbij er iets wordt toegevoegd of weggelaten.

- Aantal rotaties (F_7): De figuur moet geroteerd worden in een bepaalde richting en met een bepaalde hoek om tot de juiste oplossing te komen.

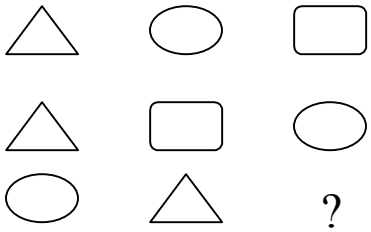


- Aantal ontbrekende gegevens in item (F_8): In sommige items ontbreken er naast het antwoord ook nog een aantal andere elementen van de matrix.

Vb: 

antwoord 

- Het aantal elementen van een bepaald type in de volledige matrix is constant. (F_9): Deze bewerking kan enkel toegepast worden op items met drie reeksen. In de 3*3 matrix moeten 3 verschillende geometrische figuren 3 keer voorkomen.

Vb: 

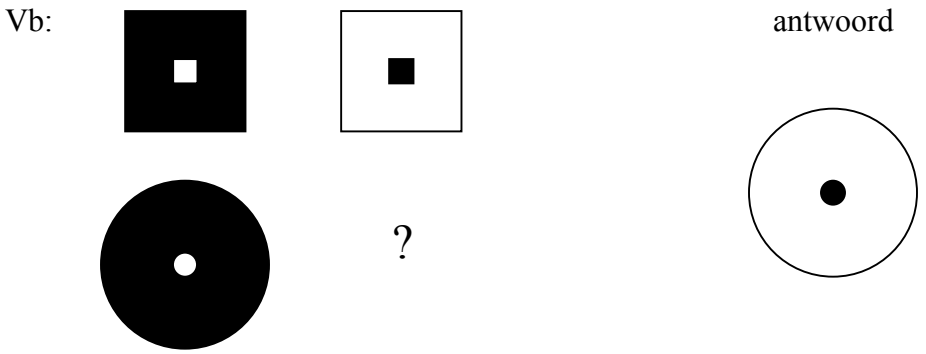
antwoord 

- Verschuiven (F_{10}=1) of niet (F_{10}=0) : In de geometrische figuur vindt er een verschuiving plaats, die moet voortgezet worden om tot de juiste oplossing te komen

Vb: o_____ _o____ ____o_ ?

Antwoord: _____o

- Omkering figuur – achtergrond (F_{11}=1) of niet (F_{11}=0): Wat in figuur A achtergrond is, wordt in figuur B de figuur en omgekeerd.



Tabel 4.15 beschrijft het resultaat van het model waarbij de 38 ξ_i parameters gemodelleerd worden als een lineaire combinatie van de 11 itemkenmerken. De geschatte verdeling van θ voor mannen in dit model is $\theta \sim N(-0.02, 0.99)$. De gemiddelde vaardigheid van mannen verschilt net zoals bij de originele DIF analyse niet significant van die van vrouwen (p>05).

Tabel 4.15 Samenhang van DIF in moeilijkheidsgraden en itemkenmerken bij TNV

Kenmerk	Omschrijving	η	p
F_1	a) 1 reeks	-.05	.34
	b) 2 reeksen	.04	.59
	c) 3 reeksen	0	
F_2	Gelijkheid	.15	.10
F_3	Spiegeling	-.21	.0005
F_4	Verkleinen/vergroten/verbreden/versmallen	.06	.45
F_5	Motiefverandering	-.13	.09
F_6	Eenvoudige bewerking	.06	.11
F_7	Roteren	-.04	.36
F_8	Ontbrekend gegeven in item	-.05	.36
F_9	Aantal elementen van bepaald type is constant	.10	.27
F_{10}	Verschuiven	.09	.06
F_{11}	Omkering figuur achtergrond	.02	.79

Uit Tabel 4.15 blijkt dat na controle voor andere itemkenmerken in het model slechts één itemkenmerk een significante samenhang vertoont met de DIF parameters. Items waarbij er gespiegeld moet worden blijken moeilijker voor vrouwen dan voor mannen met dezelfde positie op de schaal. Drie van deze items (item 14, 15 en 16) worden significant minder goed opgelost door vrouwen en de overige spiegelitems (item 6, 17, 23) neigen ook in het voordeel van de mannen (negatieve ξ's in Tabel 4.14), al vertonen ze geen significante DIF. De correlatie tussen de verschillen in de moeilijkheidsgraden van de originele DIF analyse en de DIF die voorspeld worden op basis van de itemkenmerken is .66. Dit wil zeggen dat 44% van de variantie in de DIF

parameters verklaard wordt door het model. Het grootste deel van de variantie in de DIF parameters blijft dus onverklaard.

In een tweede stap onderzoeken we in welke mate de moeilijkheidsgraden voor mannen en vrouwen kunnen verklaard worden in functie van itemkenmerken. De geschatte verdeling van θ voor mannen in dit model heeft ongeveer hetzelfde gemiddelde maar een kleinere spreiding, namelijk, θ~N(-0.02,0.77).

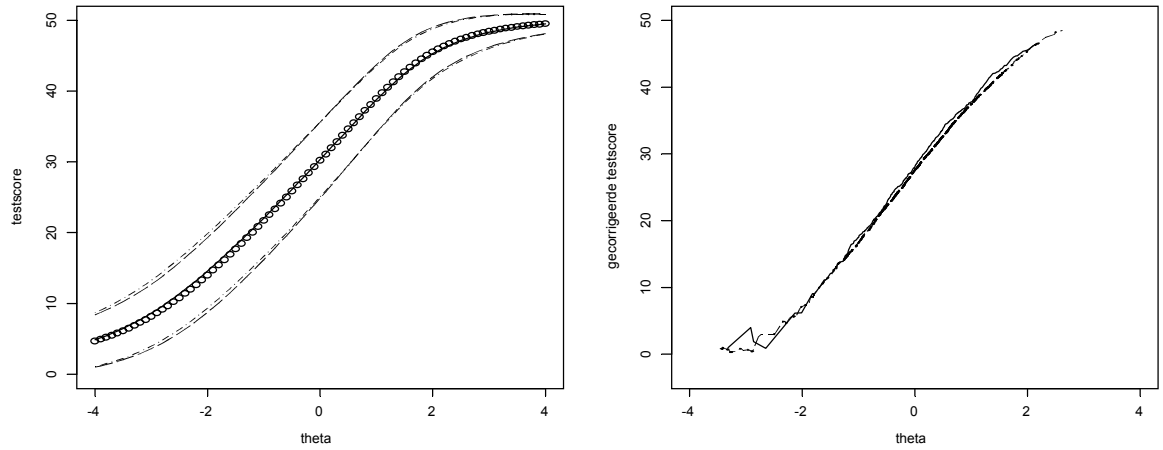
Tabel 4.16 Samenhang van moeilijkheidsgraden per groep en itemkenmerken voor TNV

Kenmerk	Omschrijving	τ	p	η	p
F_0	Constante	-1.89	<.0001		
F_1	a) 1 reeks	.71	<.0001	-.03	.54
	b) 2 reeksen	-.29	<.0001	-.01	.83
	c) 3 reeksen	0		0	
F_2	Gelijkheid	.47	<.0001	.12	.11
F_3	Spiegeling	.23	<.0001	-.16	.0014
F_4	Verkleinen/vergroten/verbreden/versmallen	.19	.0005	.02	.67
F_5	Motiefverandering	1.36	<.0001	-.06	.34
F_6	Eenvoudige bewerking	.61	<.0001	.04	.09
F_7	Roteren	.72	<.0001	-.04	.34
F_8	Ontbrekend gegeven in item	.64	<.0001	-.05	.29
F_9	Aantal elementen van bep. type is constant	1.15	<.0001	.1	.22
F_{10}	Verschuiven	1.24	<.0001	.08	.1
F_{11}	Omkering figuur achtergrond	2.4	<.0001	.22	.02

De positieve τ gewichten voor F_2 t.e.m. F_{11} geven aan dat items moeilijker worden als er meer cognitieve operaties moeten uitgevoerd worden om het item op te lossen (meer spiegelingen, meer rotaties, enz...). Daarnaast blijkt uit de η gewichten in Tabel 4.16 dat sommige itemkenmerken een verschillend gewicht hebben voor mannen en vrouwen. Analoog aan de originele DIF analyse zijn items waarbij er gespiegeld moet worden relatief moeilijker voor vrouwen dan voor mannen. Verder geldt dat items waar er een omkering van de figuur en achtergrond plaatsvindt (27, 42 en 49) moeilijker zijn voor mannen dan voor vrouwen. Uit de correlaties tussen geobserveerde en voorspelde moeilijkheidsgraden blijkt dat zowel voor mannen als voor vrouwen de moeilijkheidsgraden in elke groep matig voorspeld kunnen worden op basis van de 11 itemkenmerken. Voor mannen bedraagt deze correlatie .69 en voor vrouwen .66. De itemkenmerken verklaren dus respectievelijk 48% en 44% van de variantie in de moeilijkheidsgraden van de items bij mannen en vrouwen.

4.6.3 Effect van DIF op de testscore

Figuur 4.12 toont dat de somscore-curves en de gecorrigeerde somscore-curves voor beide groepen bijna niet verschillen, wat wil zeggen dat voor geen enkele waarde van θ er een significant verschil is tussen somscores of gecorrigeerde somscores van mannen en vrouwen op de test.



Figuur 4.12 Verwachte testscore voor mannen (-) en vrouwen (o) en 95%-betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ ._) (linkerpaneel); Verwachte gecorrigeerde testscore voor mannen (_ _) en vrouwen (__) (rechterpaneel)

4.6.4 Conclusie

We kunnen besluiten dat mannen en vrouwen gemiddeld even goed presteren op deze test voor algemene intelligentie. Een beperkt aantal items (9 van de 50) vertoont uniforme DIF met een beperkte praktische significantie. De grootste verschillen in succeskansen variëren van .10 tot .15. De vastgestelde DIF kan in beperkte mate verklaard worden. Ten eerste blijkt dat items waarbij er gespiegeld moet worden relatief moeilijker zijn voor vrouwen dan voor mannen. Analyses waarbij de moeilijkheidsgraden van elke groep gemodelleerd worden, bevestigen dit resultaat en tonen dat het omkeren van figuur en achtergrond dan weer moeilijker is voor mannen dan voor vrouwen. Ten slotte blijkt dat de gezamenlijke invloed van DIF in individuele items geen effect heeft op (verwachte) somscores en gecorrigeerde somscores.

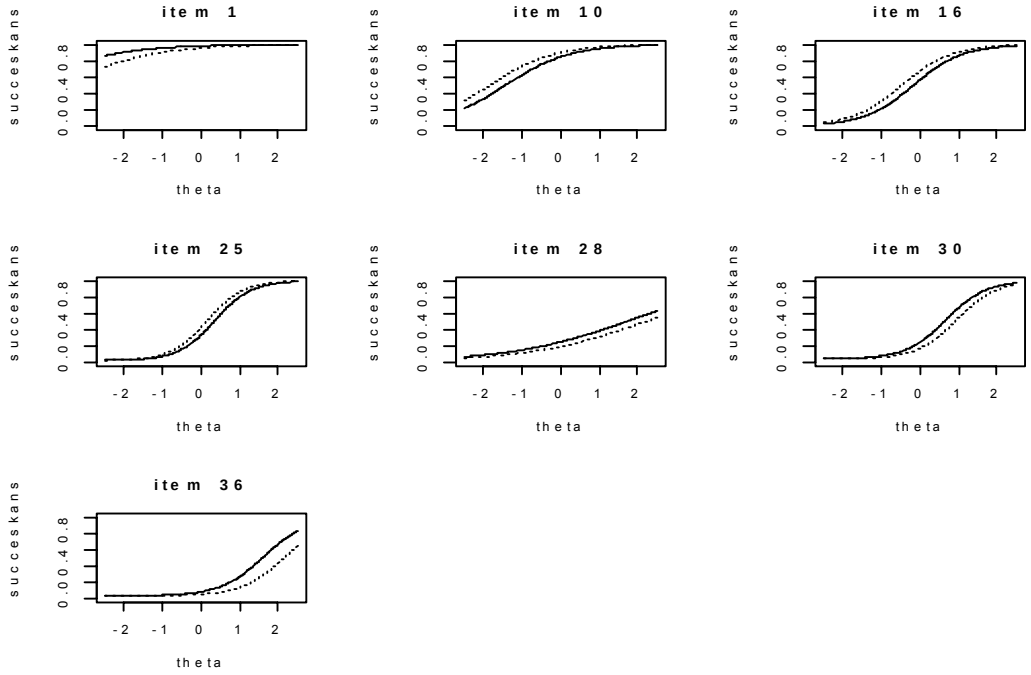
4.7 DGEO

4.7.1 Modelleren van DIF

Op basis van de methode met gelijke populatie gemiddelden werd een anker van 9 items gekozen. Deze items worden vet weergegeven in Tabel 4.17. Vervolgens wordt in SAS het model met uniforme en het model met niet-uniforme DIF voor alle niet-anker items geschat. Items 37 tot 40 worden niet opgenomen in de analyse omdat de parameters niet betrouwbaar kunnen geschat worden (grote standaardfouten).Vergelijking van de fit van modellen met uniforme en met niet-uniforme DIF aan de hand van een LR-toets toont dat de nulhypothese van gelijke discriminatiegraden in beide groepen (uniforme DIF) niet kan verworpen worden (LR=32, df=27 p=.23).

De latente variabele θ wordt voor vrouwen vastgelegd (θ~N(0,1)) en voor mannen wordt het gemiddelde en de spreiding van θ geschat uit de gegevens. Mannen scoren gemiddeld beter dan vrouwen op de test (μ=.37, p<.0001). De spreiding van θ bij mannen (σ^2=1.05) is ongeveer dezelfde als de spreiding van θ bij vrouwen

Uit Tabel 4.17 blijkt dat 13 van de 27 niet-anker items niet-uniforme DIF vertonen. De praktische significantie van de DIF is eerder beperkt: de mediaan van de absolute verschillen tussen IRFs is steeds kleiner dan .05 en voor items met significante DIF is het 97.5 percentiel van steeds kleiner dan .15. Een uitzondering is item 36 waar het verschil in succeskansen kan oplopen tot .23. Figuur 4.13 toont de IRFs voor items die significante uniforme DIF vertonen.

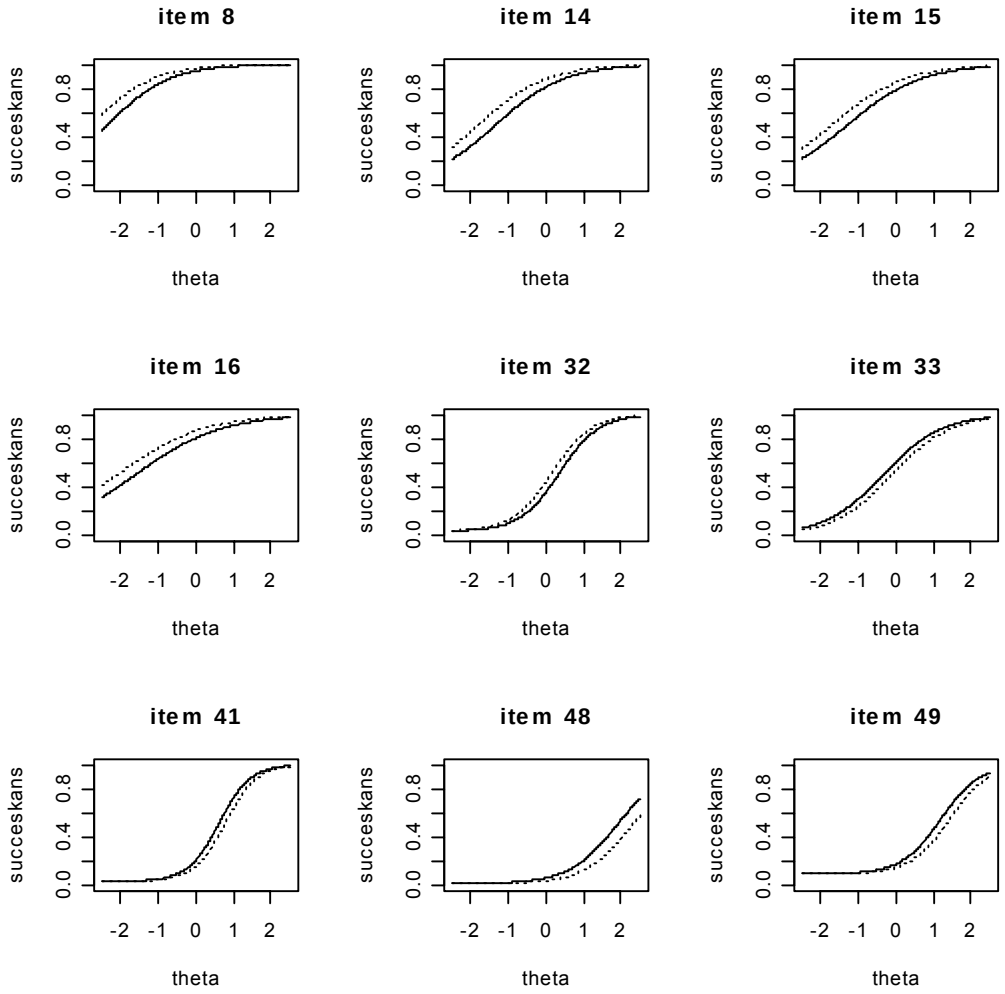


Figuur 4.13 IRFs van mannen (- -) en vrouwen (_) voor items die significante uniforme DIF vertonen (p<.01)

Tabel 4.17 Parameters van uniforme DIF analyse. Mediaan en 95% BI van absolute verschillen tussen IRFs voor mannen en vrouwen.

item	γ	α	β	ξ	Aantal keer plooien	rond circel plooien	Mediaan abs. verschil IRFs	95% BI
1	.03	.88	-4.55	.99**	0	**1**	.02	[.00,.13]
2	.03	1.12	-2.74	.16	3	1	.01	[.00,.04]
3	.03	1.2	-2.48	.41*	5	0	.03	[.00,.12]
4	.03	1.6	-2.49	0	/	/	/	/
5	.04	1.37	-1.75	-.17	6	0	.02	[.00,.06]
6	.15	1.21	.26	0	/	/	/	/
7	.02	1.16	-1.17	-.2	0	**1**	.03	[.00,.06]
8	.03	.76	-.71	0	/	/	/	/
9	.01	.86	-.84	-.28*	4	0	.05	[.01,.06]
10	.02	1.22	-1.41	-.41**	4	0	.05	[.00,.12]
11	.01	.78	-.89	0	/	/	/	/
12	.01	1.84	-1.6	-.02	8	0	.00	[.00,.01]
13	.01	1.17	-.42	.23*	4	0	.04	[.01,.07]
14	.01	1.7	.11	-.09	8	0	.01	[.00,.04]
15	.01	1.53	-.71	-.1	0	**1**	.02	[.00,.04]
16	.01	1.57	-.17	-.30**	4	0	.05	[.01,.12]
17	.01	1.01	.33	.26*	3	1	.04	[.01,.06]
18	.01	1.11	-.96	0	/	/	/	/
19	0	1.71	.31	0	/	/	/	/
20	.01	1.76	-.68	-.17	8	0	.03	[.00,.07]
21	.01	1.22	-.26	0	/	/	/	/
22	.01	.98	-.06	-.07	5	0	.01	[.01,.02]
23	.02	.66	1.08	.40*	0	**1**	.05	[.02,.06]
24	.04	1.4	.42	-.14	7	0	.02	[.00,.05]
25	.03	2.13	.35	-.19**	6	0	.02	[.00,.10]
26	.05	1.96	.68	-.13	5	0	.02	[.00,.06]
27	.04	1.86	.02	.11	5	0	.02	[.00,.05]
28	.02	.66	1.74	.51**	0	**1**	.05	[.02,.08]
29	.01	1.02	.7	0	/	/	/	/
30	.05	1.95	.67	.28**	5	0	.04	[.00,.13]
31	.01	2.62	1.34	-.07	6	0	.01	[.00,.05]
32	.03	2.25	1.38	.1	4	0	.01	[.00,.05]
33	.03	1.62	2.55	0	/	/	/	/
34	.04	1.84	1.93	.34*	0	**1**	.01	[.00,.15]
35	.01	2.36	1.89	.25	8	0	.01	[.00,.14]
36	.04	1.77	1.63	.56**	0	**1**	.03	[.00,.23]

*p<.05; **p<.01; ankeritems worden vet weergegeven

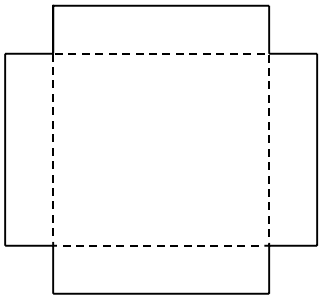


4.7.2 Verklaren van DIF

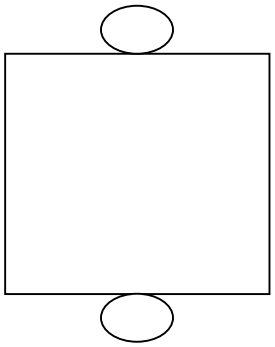
Voor het verklaren van DIF onderzoeken we of de geschatte DIF parameters ξ kunnen verklaard worden in functie twee itemkenmerken:

- F_1: Het aantal keren dat er mentaal geplooid moet worden in een item. Bijvoorbeeld, in item A moet er 4 maal mentaal geplooid worden. (Zie Tabel 4.17: aantal keer plooien)
- F_2: Of men in een cirkel moet plooien (F_2=1) of niet (F_2=0). Bijvoorbeeld, in item B moet men de zijden van een rechthoek cirkelvormig plooien om een cilinder te vormen. (Tabel 4.17: rond cirkel plooien=1, niet rond circel plooien=0)

itemA



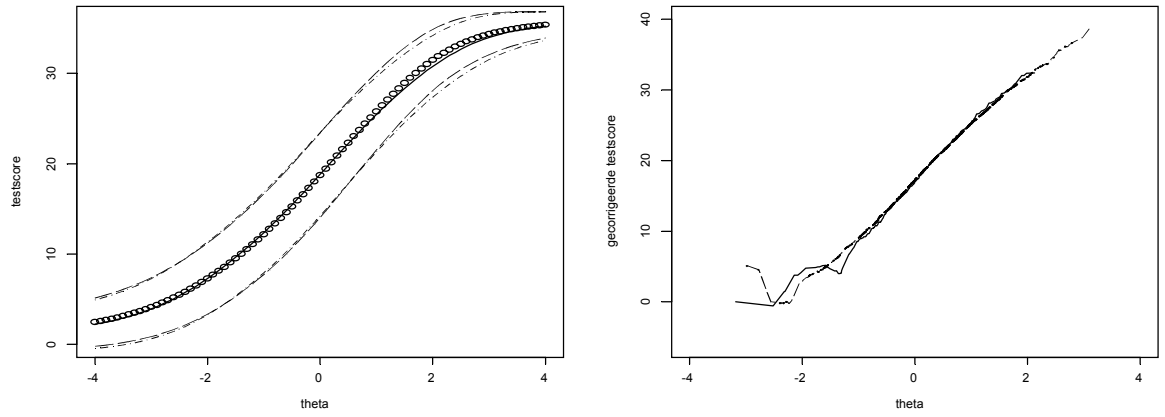
item B



Uit een cognitieve analyse van de items volgt dat F_1 en F_2 hoog negatief gecorreleerd zijn (r = -0.86), zodat hun gezamenlijke invloed op de DIF parameters niet betrouwbaar kan onderzocht worden. Wanneer we de afzonderlijke correlaties van F_1 en F_2 met het verschil in moeilijkheidsgraad tussen mannen en vrouwen (ξ) bekijken, zien we dat F_1 significant negatief correleert met ξ (r = -.48, p=.011) terwijl F_2 significant positief correleert met ξ (r = .52, p=.006) Dit wil zeggen dat items waar meer mentaal geplooid moet worden, moeilijker zijn voor vrouwen dan voor mannen en dat items waar rond een cirkel moet geplooid worden, moeilijker zijn voor mannen dan voor vrouwen. In tabel 4.17 wordt per item het aantal mentale plooien weergegeven. Tussen de moeilijkheidsgraden van de items en de beide itemkenmerken zijn er echter geen significante verbanden. De moeilijkheidsgraden van vrouwen correleren respectievelijk .01 en .07 met F_1 en F_2 en de moeilijkheidsgraden van de mannen correleren respectievelijk -.08 en .17 met F_1 en F_2.

4.7.3 Effect van DIF op de testscore

Figuur 4.14 toont dat er voor geen enkele waarde van θ een significant verschil is tussen de (verwachte) somscores en gecorrigeerde somscores van mannen en vrouwen.



Figuur 4.14 Verwachte testscore voor mannen (-) en vrouwen (o) en 95%-betrouwbaarheidsinterval voor mannen (_ _) en vrouwen (_ ._) (linkerpaneel); Verwachte gecorrigeerde testscore voor mannen (_ _) en vrouwen (__) (rechterpaneel)

4.7.4 Conclusie

We stellen vast dat mannen gemiddeld beter presteren op deze test voor ruimtelijk inzicht dan vrouwen. Bij één op drie items treedt uniforme DIF op met een beperkte praktische significantie. De mediaan van de absolute verschillen tussen IRFs van mannen en vrouwen is steeds kleiner dan .05 en voor items die statistisch significante DIF vertonen, is het 97.5 percentiel van de absolute verschillen zelden groter dan .15. De DIF kan in beperkte mate verklaard worden op basis van twee itemkenmerken, namelijk het aantal keer dat men mentaal moet plooien en of er recht of cirkelvormig moet geplooid worden. Naarmate er meer mentaal moet geplooid worden, worden de items moeilijker voor vrouwen en cirkelvormig mentaal plooien blijkt moeilijker voor mannen. Tenslotte stellen we vast dat de gezamenlijke invloed van DIF in individuele items niet leidt tot verschillende (verwachte) somscores en gecorrigeerde somscores voor mannen en vrouwen.

Hoofdstuk 5. Samenvatting van onderzoeksresultaten en beleidsaanbevelingen

5.1 Samenvatting van onderzoeksresultaten

In dit rapport werd voor zeven verschillende types van intelligentietests die frequent gebruikt worden voor personeelsselectie bij SELOR en ABL onderzocht of ze discriminerend zijn voor vrouwen versus mannen. 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 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.

5.1.1 Gemiddelde prestaties van mannen en vrouwen

Tabel 5.1 geeft voor de onderzochte tests een overzicht van de voornaamste testkarakteristieken (grootte van de data set, interne consistentie) en van de gemiddelde prestaties van mannen en vrouwen. De interne consistentie van de tests (bepaald aan de hand van coëfficiënt α) van ABL is hoog tot zeer hoog (>.85) en verschilt zeer weinig voor mannen en vrouwen. Bij de tests van SELOR is de interne consistentie alleen hoger dan .85 voor de lange tests (ANAVERB en CODES) en zijn de verschillen in interne consistentie bij mannen en vrouwen groter.

De gemiddelde prestaties van mannen en vrouwen worden weergegeven door het gemiddelde van de latente variabele θ in elke groep. Voor vrouwen werd de gemiddelde θ waarde vastgelegd op 0. Voor mannen wordt de gemiddelde θ waarde bepaald op basis van de gegevens. Uit Tabel 5.1 blijkt dat mannen gemiddeld beter presteren dan vrouwen op vier van de zeven tests.
Het gemiddelde van de vaardigheidsverdeling van mannen is significant hoger dan het gemiddelde van de vrouwen voor LOGDED (μ=.27), ANAVERB (μ=.42), WIMA (μ=.27) en DGEO (μ=.40). Voor de andere drie tests (CODES, NUMVA, TNV) is er geen significant verschil tussen de gemiddelde vaardigheden van mannen en vrouwen.

Om de praktische significantie van deze effecten na te gaan kunnen we berekenen hoeveel de succeskans toeneemt voor een bepaald item als het opgelost wordt door een man met gemiddelde vaardigheid in plaats van door een vrouw met gemiddelde vaardigheid. Veronderstel bijvoorbeeld zuivere items (ξ_i=0 en ε_i=0) met lage, gemiddelde of hoge discriminatiegraad (α_i=0.5, 1 en 2, respectievelijk), met moeilijkheidsgraad (β_i) gelijk aan 0 en met raadparameter (γ_i) gelijk aan 0. Op basis van formule (3.3) kan men eenvoudig berekenen dat voor dit soort items de succeskans van vrouwen met een gemiddelde vaardigheid (θ=0) gelijk is aan .50. Mannen met een gemiddelde vaardigheid hebben voor dit soort items in het algemeen een succeskans hoger dan .50. Zoals blijkt uit Tabel 5.1 wordt de grootte van dit effect sterk bepaald door de discriminatieparameter van het item. Bijvoorbeeld, voor de tests met maximale verschillen in

gemiddelde vaardigheid (ANAVERB en DGEO) zien we dat voor zwak discriminerende items (α_i=0.5) de succeskans van de gemiddelde man slechts .05 hoger is dan die van de gemiddelde vrouw. Bij items met een matige discriminatiegraad (α_i=1) lopen de verschillen in succeskansen op tot .10, en bij items met een hoge discriminatiegraad kunnen de verschillen in succeskansen zelfs oplopen tot .20.

Men zou op basis van de resultaten in Tabel 5.1 kunnen geneigd zijn te besluiten dat mannen gemiddeld beter presteren op tests die een bepaald soort intelligentie meten. Maar deze conclusie is alleen gerechtvaardigd als deze tests door gelijkaardige steekproeven van mannen en vrouwen werden opgelost of als de overlap tussen de verschillende tests groot genoeg is om de gegevens van alle tests tegelijk te modelleren. De reden waarom men op basis van verschillende gemiddelde prestaties van verschillende steekproeven op verschillende types van tests geen conclusies kan trekken is omdat het zou kunnen dat bij de éne test de steekproef van mannen toevallig intelligenter is (bijvoorbeeld omdat ze gemiddeld een hoger opleidingsniveau hebben) en dat bij een andere test toevallig de vrouwen intelligenter zijn. Rekening houdend met deze bedenking kan men op basis van de resultaten van ABL sterkere conclusies trekken dan op basis van de resultaten van SELOR.

De steekproeven van de drie ABL tests overlappen sterk omdat deze tests werden aangeboden aan alle kandidaten die in een defensiehuis kwamen informeren voor een functie bij het leger. Dit onderzoek toont dus aan dat mannen en vrouwen gemiddeld even goed presteren op de algemene intelligentie test TNV maar dat mannen gemiddeld beter presteren op de vraagstukken test WIMA en op de ruimtelijk inzicht test DGEO. Deze resultaten zijn in overeenstemming met wat in intelligentie-onderzoek al eerder werd vastgesteld: Enerzijds werd vastgesteld dat geslachtsverschillen in algemene intelligentie verwaarloosbaar zijn (Jensen, 1998; Halpern, 2000; Fabregat, Colomb, Abad, Espinosa, 2000). Anderzijds werd in de literatuur vastgesteld dat mannen beter zijn dan vrouwen in taken waarbij visuo-spatiale vaardigheden (spatiale perceptie, mentale rotatie, spatiale visualisatie, spatiotemporele vaardigheden, genereren en onthouden van spatiale informatie) een rol spelen (Casey, Nuttall & Pezaris, 1997). Visuo-spatiale vaardigheden zijn cruciaal bij het oplossen van de DGEO test en kunnen soms ook helpen bij het oplossen van vraagstukken als hiervoor een visuele strategie gevolgd wordt.

De steekproeven van de SELOR tests vertonen maar een beperkte overlap zodat we op basis van de bevindingen van ons onderzoek in principe geen sterke conclusies kunnen trekken over het gemiddeld presteren van mannen en vrouwen op de verschillende tests. Het zou bijvoorbeeld kunnen dat in de steekproef die LOGDED oploste toevallig meer mannen zitten uit exact wetenschappelijke richtingen en dat deze mannen beter getraind zijn in logisch redeneren. Om dit soort verklaring te valideren is het noodzakelijk dat selecteurs informatie over mogelijke verklarende factoren (zoals opleidingsniveau) systematisch registreren en inventariseren samen met de ruwe testgegevens. Hoewel de resultaten van SELOR i.v.m. het gemiddeld presteren van mannen en vrouwen ons niet toelaten om harde uitspraken te doen dienen we hier toch op te merken dat ze niet geheel onwaarschijnlijk zijn omdat ze in overeenstemming zijn met de literatuur van het intelligentie onderzoek. Zo werd bijvoorbeeld reeds vastgesteld dat mannen beter zijn in verbale-analogie taken (zoals ANAVERB) (Halpern & Wright, 1996) en in taken waar visuele strategieën een rol kunnen spelen (zoals de LOGDED test).

Tabel 5.1 Overzicht van onderzochte tests bij SELOR en ABL. Grootte van de data set, interne consistentie en gemiddelde vaardigheid van mannen en vrouwen.

Organisatie	Test	Aantal items	Aantal mannen	Interne Consistentie mannen	Aantal vrouwen	Interne Consistentie vrouwen	Gemiddelde θ mannen	Succeskans man met gemiddelde vaardigheid		
								α_i=.5	α_i=1	α_i=2
SELOR	LOGDED	22	1895	.80	2526	.73	.27**	.53	.57	.63
SELOR	ANAVERB	98	1656	.99	2401	.90	.42**	.55	.60	.70
SELOR	CODES	74	731	.96	574	.95	-.09	.49	.48	.46
SELOR	NUMVA	38	1113	.84	1227	.80	.10	.51	.52	.55
ABL	WIMA	23	2257	.87	341	.86	.27**	.53	.57	.63
ABL	TNV	50	2963	.87	431	.89	.08	.51	.52	.54
ABL	DGEO	36	2962	.89	431	.88	.40**	.55	.60	.69

Tabel 5.2 Proportie items per test die DIF vertonen volgens verschillende criteria

Test	Aantal mannen	Aantal vrouwen	Proportie items met DIF op 5% niveau	Proportie van alle items met bepaalde waarde voor Mediaan van de absolute verschillen tussen IRFs van mannen en vrouwen			Proportie van alle items met bepaalde waarde voor 97.5 percentiel van de absolute verschillen tussen IRFs van mannen en vrouwen			
				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
LOGDED	1895	2526	.27	.95	.05	.00	.86	.09	.05	.00
ANAVERB	1656	2401	.51	.87	.10	.03	.52	.22	.16	.10
CODES	731	574	.07	1.00	.00	.00	.73	.23	.03	.01
NUMVA	1113	1227	.29	.89	.11	.00	.55	.18	.16	.11
WIMA	2257	341	.30	.78	.22	.00	.74	.22	.04	.00
TNV	2963	431	.18	.94	.06	.00	.86	.12	.02	.00
DGEO	2962	431	.36	.86	.14	.00	.75	.19	.03	.03

Dat mannen (vrouwen) gemiddeld beter presteren dan vrouwen (mannen) op een bepaalde intelligentie test heeft als gevolg dat er systematisch meer mannen geselecteerd zullen worden op basis van deze test. Dit is in principe gerechtvaardigd als de test het succesvol uitoefenen van een bepaalde functie beter voorspelt dan andere tests die wel even moeilijk zijn voor mannen en vrouwen. Men dient er evenwel over te waken dat de tests die gebruikt worden voor een bepaalde selectie afgestemd zijn op de specifieke vaardigheden die nodig zijn voor het succesvol uitoefenen van een bepaalde job.

5.1.2 DIF in moeilijkheidsgraden

Zoals blijkt uit Tabel 5.2 treedt er bij alle onderzochte tests voor een aantal items DIF op in de moeilijkheidsgraden. De proportie van alle items die significante DIF vertonen op het 5% niveau varieert over tests van .07 tot .51. De meeste tests vertonen DIF in ongeveer één derde van de items. Bij de CODES test is er DIF in minder dan 10% van de items en bij ANAVERB is er DIF in ongeveer de helft van de items.

Omdat de statistische significantie van de DIF ook 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 5.2 blijkt dat de praktische significantie van de DIF voor de meeste tests eerder beperkt is. De mediaan van de absolute verschillen tussen IRFs van mannen en vrouwen is voor het grootste deel van de items kleiner dan .05. Verder geldt dat in bepaalde tests voor een beperkt aantal items en voor een beperkt deel van de latente schaal de DIF sterk kan oplopen. Bijvoorbeeld voor ANAVERB en NUMVA is voor respectievelijk 10% en 11% van de items het 97.5 percentiel van absolute verschillen tussen IRFs van mannen en vrouwen groter dan .20. Met andere woorden, in ongeveer 10% van de ANAVERB en NUMVA items zijn de verschillen tussen succeskansen van mannen en vrouwen op een klein stuk van de latente schaal (2.5 %) minstens .20. Op basis van de resultaten in de Tabel kunnen we besluiten dat de DIF eerder beperkt is en dat alleen bij NUMVA en ANAVERB het interessant zou zijn om een beperkt aantal items te verwijderen om de constructvaliditeit van de test te verbeteren.

5.1.3 Gezamenlijk effect van DIF in individuele items op somscores en gecorrigeerde 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 en gecorrigeerde somscores. Met andere woorden, het verband tussen de latente variabele en de prestatie op de ganse test is voor de onderzochte tests hetzelfde in beide groepen. Alleen voor ANAVERB is het interessant om een beperkt aantal items te verwijderen die sterke DIF in het voordeel van de mannen vertonen. Bij ANAVERB blijkt immers dat mannen met een lage vaardigheid een aanzienlijk hogere verwachte somscore hebben dan vrouwen (hoewel niet significant). Men kan dus stellen dat somscores en gecorrigeerde somscores valide maten zijn om kandidaten onderling of uit verschillende groepen te vergelijken.

Maar men moet er zich wel van bewust zijn dat eenzelfde testscore iets anders kan betekenen voor mannen en vrouwen omdat ze het resultaat kan zijn van succes op verschillende verzamelingen van items die elk een tegengestelde DIF vertonen.

Voor sommige onderzochte tests blijkt dat het 95% betrouwbaarheidsinterval dat kan afgebakend worden rond de somscores erg breed is. Bijvoorbeeld, in de steekproef van vrouwen bij LOGDED is de standaardmeetfout uit de klassieke testtheorie gelijk aan 1.77 wat wil zeggen dat een somscore 12 met 95% zekerheid ligt tussen 12-(1.96*1.77)=8.5 en 12+(1.96*1.77)=15.5. Bij het ordenen van kandidaten op basis van de somscore is het van belang om hiermee rekening te houden. Kandidaten kunnen immers strikt genomen slechts 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.

5.1.4 Verklaren van DIF

Aangezien de onderzochte tests allemaal voor bepaalde items DIF vertonen is het belangrijk om te onderzoeken op welke manier de DIF verklaard kan worden. Een inzicht in welke items relatief moeilijker zijn voor mannen of vrouwen geeft immers een beter inzicht in de constructvaliditeit van de test en in de betekenis van de somscores van mannen en vrouwen.

Voor elk van de onderzochte tests werd nagegaan in welke mate DIF in de moeilijkheidsgraden kan verklaard worden op basis van objectieve itemkenmerken. Daarnaast werd ook onderzocht in welke mate de moeilijkheidsgraden in elke groep kunnen verklaard worden op basis van itemkenmerken. De itemkenmerken werden telkens bepaald op basis van een cognitieve analyse van de testitems.

Tabel 5.3. Percentage van de variantie in DIF parameters (ξ) dat kan verklaard worden op basis van itemkenmerken. Percentage van de variantie in moeilijkheidsgraden voor mannen en vrouwen dat kan verklaard worden op basis van itemkenmerken.

Test	Aantal gemodelleerde parameters	Aantal item kenmerken	Percentage verklaarde variantie		
			Dif-parameter	Moeilijkheidsgraad	
				Vrouwen	mannen
LOGDED	17	5	23	77	77
ANAVERB	98	7	7	14	13
CODES	60	5	15	35	34
NUMVA	38	6	22	40	41
WIMA	13	8	64	34	46
TNV	38	11	44	44	48
DGEO	27	2	25	1	1

Tabel 5.3 toont voor elke test hoe goed (in termen van percentage verklaarde variantie) DIF in de moeilijkheidsgraden of moeilijkheidsgraden voor mannen en vrouwen zelf kan verklaard worden op basis van itemkenmerken. We stellen vast dat het voor de meeste tests erg moeilijk is om te verklaren waarom bepaalde items een verschillende moeilijkheidsgraad hebben voor mannen en vrouwen. Bij de meeste tests blijft het grootste deel van de variantie onverklaard. Alleen bij WIMA blijkt dat meer dan de helft (64%) van de variantie in de DIF parameters kan verklaard worden op basis van itemkenmerken. We moeten hier echter bij opmerken dat bij WIMA slechts 13 parameters (DIF in niet-anker items) gemodelleerd worden op basis van 8 itemkenmerken. Het is daarom erg waarschijnlijk dat het model een deel van de ruis in de moeilijkheidsgraden modelleert en dat het model maar beperkt de moeilijkheidsgraden van nieuwe WIMA items kan voorspellen.

De resultaten voor het verklaren van de moeilijkheidsgraden voor mannen en vrouwen zijn analoog. Voor de meeste tests blijft het grootste deel van de moeilijkheidsgraden onverklaard. Alleen voor LOGDED kunnen de moeilijkheidsgraden redelijk goed gevat worden in termen van een beperkt aantal cognitieve operaties die nodig zijn om de items op te lossen.

We kunnen besluiten dat het modelleren van moeilijkheidsgraden van testitems op basis van een inhoudelijke theorie over de cognitieve processen die een rol spelen bij het oplossen van de items, of op basis van andere objectieve itemkenmerken slechts een goede kans op slagen heeft als de test geconstrueerd werd op basis van een theorie en als een beperkt aantal itemkenmerken of processen in voldoende items herhaaldelijk aan bod komen (zoals bij de LOGDED test). Als een test relatief veel verschillende itemkenmerken bevat die telkens slechts bij enkele items aanbod komen dan is het veel moeilijker om te begrijpen wat een item moeilijk maakt en dan is het erg waarschijnlijk dat de resultaten slechts beperkt generaliseerbaar zijn naar nieuwe testitems.

Hoewel de DIF in de moeilijkheidsgraad maar beperkt verklaard kan worden, toch werden voor verschillende tests gelijkaardige verbanden gevonden tussen itemkenmerken en de geobserveerde DIF. We geven een overzicht van enkele bevindingen die van toepassing zijn op verschillende tests.

1. Bij LOGDED stellen we vast dat items met ≤ in de propositie van het antwoordalternatief gemakkelijker zijn voor mannen. Een mogelijke verklaring hiervoor is dat deze items gemakkelijker zijn als men een visuele strategie gebruikt en dat mannen sneller geneigd zijn om deze strategie te gaan gebruiken. Een analoge bevinding die mogelijks dezelfde verklaring heeft is dat TNV-items waarbij moet gespiegeld worden gemakkelijker zijn voor mannen.

2. Bij LOGDED blijkt dat items met meer premissen gemakkelijker zijn voor vrouwen. Een analoge bevinding bij WIMA is dat items die meer stappen vereisen gemakkelijker zijn voor vrouwen.

Deze bevindingen geven een eerste aanwijzing van hoe de onderzochte tests mogelijks in lichte mate verschillende constructen meten voor mannen en vrouwen. Er is echter verder onderzoek nodig om te evalueren in welke mate de bevindingen van dit rapport (verschillen in) moeilijkheidsgraden van nieuw geconstrueerde items kunnen voorspellen.

5.2 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.

5.2.1 Advies voor testproducenten en selecteurs

1. Het onderzoek wijst uit dat vrouwen op bepaalde tests gemiddeld slechter scoren dan mannen (logisch redeneren, verbale analogieën, mathematische vraagstukken, ruimtelijk inzicht) terwijl er voor andere tests geen verschillen zijn tussen mannen en vrouwen (algemene intelligentie, cijferreeksen, leren van een code). Als mannen (vrouwen) gemiddeld beter presteren op een bepaalde test dan wil dit zeggen dat zij in principe meer kans hebben om geselecteerd te worden. De groep die gemiddeld lager scoort wordt dan gediscrimineerd. Deze discriminatie is evenwel terecht als de specifieke vaardigheden die gemeten worden door de test ook werkelijk van belang zijn voor de succesvolle uitoefening van de job. Als het daarentegen niet duidelijk is dat deze specifieke vaardigheden leiden tot een betere jobperformantie dan is de discriminatie onterecht en dan zou men beter tests gebruiken die even moeilijk zijn voor mannen en vrouwen.

Om onterechte discriminatie van mannen of vrouwen in de selectiepraktijk te vermijden is het nodig om voor alle tests zo goed mogelijk in te schatten of ze moeilijker zijn voor mannen of voor vrouwen. Tests die niet even moeilijk zijn voor mannen en vrouwen zouden alleen mogen gebruikt worden als aangetoond is dat de vaardigheden die ze meten echt noodzakelijk zijn voor succes in de job. Anders geeft men beter de voorkeur aan tests die even moeilijk zijn voor beide groepen.

Het is dus van belang om voor alle tests die gebruikt worden zo goed mogelijk de gemiddelde vaardigheid van mannen en vrouwen in te schatten. De resultaten van ons onderzoek geven al een eerste aanwijzing over het verschil in de gemiddelde vaardigheid van mannen en vrouwen voor de onderzochte tests. Verder onderzoek is echter nodig om na te gaan of het vastgestelde verschil niet gedeeltelijk kan verklaard worden door kenmerken van de onderzochte steekproeven, zoals bijvoorbeeld het opleidingsniveau van de mannelijke en vrouwelijke kandidaten. Het zou bijvoorbeeld kunnen dat toevallig meer mannen in de steekproef een exact wetenschappelijke opleiding hebben wat de betere prestaties voor logisch redeneren gedeeltelijk zou verklaren. Daarnaast is het nodig om ook voor andere veel gebruikte tests te onderzoeken of ze moeilijker zijn voor mannen of voor vrouwen. Om zo goed mogelijk in te schatten of tests moeilijker zijn voor een bepaalde groep is het nodig om tijdens selecties de (anonieme) ruwe gegevens en bepaalde achtergrondvariabelen systematisch te registreren, wat op dit moment nog niet het geval is.

Om onterechte discriminaties te vermijden is het nodig om voor alle tests die gebruikt worden voor selectie zo goed mogelijk in te schatten of ze moeilijker zijn voor een bepaalde groep. Tests die moeilijker zijn voor een bepaalde groep zouden alleen gebruikt mogen worden voor selectie als aangetoond is dat de vaardigheden die ze meten sterker samenhangen met succes in de job dan andere vaardigheden waar mannen en vrouwen gemiddeld even goed in zijn.

Om in te schatten of tests moeilijker zijn voor een bepaalde groep dient men tijdens selecties (met het nodige respect voor de privacy) de ruwe testgegevens en relevante achtergrondvariabelen systematisch te registreren. Dit laat toe via de gepaste analyses de kwaliteit van tests te bewaken.

2. De resultaten van het onderzoek tonen aan dat alle onderzochte tests een aantal "discriminerende" (DIF) items bevatten die moeilijker zijn voor mannen of vrouwen met dezelfde totaalscore op de test. Afhankelijk van de test varieert het aantal discriminerende items van 10% tot 50%. We kunnen stellen dat bij de meeste discriminerende items de succeskansen van mannen en vrouwen niet veel verschillen maar bij een beperkt aantal items is het verschil in succeskansen wel van betekenis. Bij de SELOR tests NUMVA (cijferreeksen) en ANAVERB (verbale analogieën) is het aangeraden om een aantal sterk discriminerende items (met locale verschillen in succeskansen die groter dan .20 kunnen zijn) te verwijderen. Omdat discriminatie bij de meeste tests blijkt voor te komen is het aangeraden om ook voor andere tests discriminatie te onderzoeken. Net zoals voor het onderzoek naar gemiddelde prestaties is het hiervoor nodig dat de ruwe testgegevens en

bepaalde persoonsgebonden variabelen (vb, geslacht) systematisch geregistreerd worden tijdens selecties.

Aangezien alle onderzochte tests een aantal discriminerende items bevatten is het aangeraden om ook voor nog andere tests te onderzoeken welke items discriminerend zijn voor mannen of vrouwen. Hiervoor is het nodig dat de ruwe testgegevens en bepaalde persoonsvariabelen geregistreerd worden tijdens de selectie.

3. Het onderzoek toont aan dat de gezamenlijke invloed van discriminatie in individuele items bij de meeste tests geen verschillend effect heeft op de testscores van mannen en vrouwen. Een uitzondering is de ANAVERB test waar mannen met dezelfde lage vaardigheid hogere testscores behalen dan vrouwen met dezelfde lage vaardigheid. Hoewel dit verschil in testscores beperkt blijft, is het toch aanbevolen om bij deze test sterk discriminerende items te verwijderen. Men moet zich er bovendien van bewust zijn dat de betekenis van de testscore anders is voor mannen en vrouwen als mannen en vrouwen verschillende succeskansen hebben op bepaalde items.

We stellen 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 (al dan niet voor raden gecorrigeerde) testscores gebruikt worden om kandidaten te vergelijken. Men moet zich er evenwel van bewust zijn dat testscores bij mannen en vrouwen een verschillende betekenis hebben. 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.

4. In sommige tests lijken vrouwen minder goed te presteren voor sommige items omdat de items een beroep doen op visuele intelligentie. Dergelijke effecten zijn onwenselijk als de test zelf niet bedoeld is om visuele intelligentie te meten.

Het is van belang om na te gaan of vrouwen geen nadeel ondervinden van items waarvoor onbedoeld een beroep wordt gedaan op visuele intelligentie. Tests zouden hierop gescreend moeten worden en de problematische items zouden moeten verwijderd worden.

5.2.2 Advies aan de overheid

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. Ons onderzoek wijst uit dat het van belang is om niet alleen aandacht te schenken aan de betrouwbaarheid en de validiteit van tests maar dat men ook dient na te gaan of een test niet moeilijker is voor een bepaalde bevolkingsgroep, en of bepaalde items niet discriminerend zijn voor een bepaalde bevolkingsgroep. In wat volgt suggereren we een aantal mogelijke maatregelen die de kwaliteit van tests in de hand kunnen werken. Deze maatregelen verschillen op vlak van haalbaarheid en strengheid.

1. 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 test producent 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 zoals voor allochtonen, enz. Het toekennen van een kwaliteitslabel zou bijvoorbeeld kunnen gebeuren door een onafhankelijke commissie van test experts (zoals bijvoorbeeld de Commissie Test Aangelegenheden Nederland die is ingesteld door het Nederlands Instituut van Psychologen).

2. 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 test experts in dienst van de overheid, of zou bijvoorbeeld kunnen uitbesteed worden aan een wetenschappelijke instelling.

Referenties

Casey, M. B., Nutall, R. L., & Pezaris, E. (1997). Mediators of gender differences in Mathematics College Entrance Test scores. *Developmental Psychology, 33,* 669-680.

Fabregat, A. A., Colomb, R., Abad, F., & Espinosa, M. J. (2000). Sex differences in general intelligence defined as g among young adolescents. *Personality and Individual Differences, 28*, 813-820.

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

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3th ed.). Mahwah, N.J.: Erlbaum.

Halpern, D. F., & Wright, T. M. (1996). A process-oriented model of cognitive sex differences. *Learning and Individual Differences, 8(1)*, 3-24.

Jensen, A.R. (1998). The g factor: *The science of mental ability*. New York: Praeger.

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.

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.

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.