9

Vraagonzuiverheid

Onzuiverheid van vragen (in het Engels 'item bias' of 'differential item functioning', afgekort DIF) en onzuiverheid van tests of toetsen ('test bias') vormen in Amerika sinds het midden van de jaren 60 een belangrijk thema in 'educational measurement'. Door een aantal juridische zaken is dit onderwerp in Amerika in de jaren 80 ook sterk in de publieke belangstelling komen te staan. Een belangrijk geval daarbij vormt de rechtszaak die verzekeringsmaatschappij Golden Rule in 1976 tegen Educational Testing Service (ETS) aanspande. De aanklacht had betrekking op de negatieve gevolgen voor kleurlingen van het gebruik van bepaalde door ETS geconstrueerde toetsen voor het diploma van verzekeringsagent. In 1984 werd tussen ETS en de betreffende verzekeringsmaatschappij een schikking getroffen. Een belangrijk punt daarin was dat voor de constructie van twee specifieke toetsen uit dit examen bij de selectie van vragen zoveel mogelijk de voorkeur zou worden gegeven aan vragen die zo klein mogelijke verschillen in moeilijkheidsgraad vertoonden tussen de meerderheidsgroep en de verschillende ethnische groepen. Daarbij zou men vooral verschillen ten nadele van minderheidsgroepen trachten te voorkomen.

In Nederland werd in 1987 naar aanleiding van verschillende klachten door het Landelijk Bureau Racismebestrijding (LBR) een onderzoeksproject 'Psychologische tests en allochtonen' gestart. Gebleken was dat een aantal allochtone sollicitanten, die gekwalificeerd waren voor een functie waarnaar zij solliciteerden, door negatieve resultaten op bepaalde psychologische tests waren afgewezen. Uit een symposium van experts dat in dat jaar georganiseerd werd, kwam de volgende aanbeveling naar voren: "Psychologische tests moeten, willen ze gehanteerd worden in een selectieprocedure, gescreend zijn op 'cultural bias' en cultuurgebonden en racistische items" (LBR, 1988). Naar aanleiding hiervan werd door de Commissie Testaangelegenheden (COTAN) van het Nederlands Instituut van Psychologen en het LBR een commissie samengesteld met als taak om de twintig meest gebruikte tests op deze punten te screenen. In 1990 volgde het rapport van deze commissie waarin twintig van de in Nederland meest gebruikte psychologische tests voor de selectie voor opleiding en beroep op deze punten werden

doorgelicht (LBR, 1990). De belangrijkste conclusie uit dit rapport was dat: "alle gescreende tests sterk beperkt toepasbaar zijn bij allochtonen" en de commissie beval voor veel van de tests een "grondige revisie aan vanwege hun ethnocentristische inhoud" aan. Verder constateerde de commissie een "ernstige achterstand in Nederland op het gebied van onderzoek naar test en item bias".

Onder andere op grond van de hierboven genoemde overwegingen wordt er op het Cito de nodige aandacht besteed aan onderzoek naar onzuiverheid. Een andere overweging is dat in verschillende onderzoeken bij examens en toetsen opvallende verschillen tussen sociale groepen en geslachtsverschillen gevonden zijn, hetgeen de vraag naar de rol van de toetsen of toetsvragen zelf daarin relevant maakt. Zo zijn er verschillende onderzoeken naar vraagonzuiverheid uitgevoerd met betrekking tot allochtonen bij de Eindtoets Basisonderwijs (Uiterwijk, 1990) en bij de eindexamens voortgezet onderwijs met betrekking tot sexe (Bügel, 1993).

Onzuiverheid van tests of vragen hoeft niet alleen betrekking te hebben op bepaalde sociale groepen maar kan ook als onderdeel van een meer algemeen probleem beschouwd worden. In het kader van het meten van leerprestaties kan men bijvoorbeeld ook de onzuiverheid van toetsen of toetsvragen ten opzichte van verschillende onderwijsmethoden beschouwen.

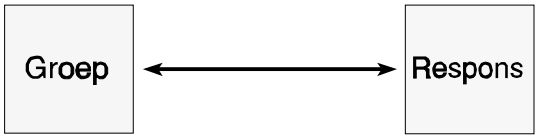
Hoewel in dit hoofdstuk ook enige aandacht aan testonzuiverheid zal worden besteed, vormt vraagonzuiverheid het belangrijkste onderwerp. In de literatuur zijn verschillende onderzoeksmethoden voor het opsporen van vraagonzuiverheid beschreven. Bij de bespreking van dergelijke methoden zullen we ons in dit hoofdstuk voornamelijk concentreren op onderzoek met behulp van IRT-modellen.

Dit hoofdstuk is als volgt opgebouwd. In paragraaf 9.1 wordt een definitie van het begrip onzuiverheid gegeven. In paragraaf 9.2 wordt deze definitie vertaald naar een aantal technieken voor het opsporen en aantonen van vraagonzuiverheid. In paragraaf 9.3 zal de toepassing van deze technieken aan de hand van een voorbeeld worden geïllustreerd.

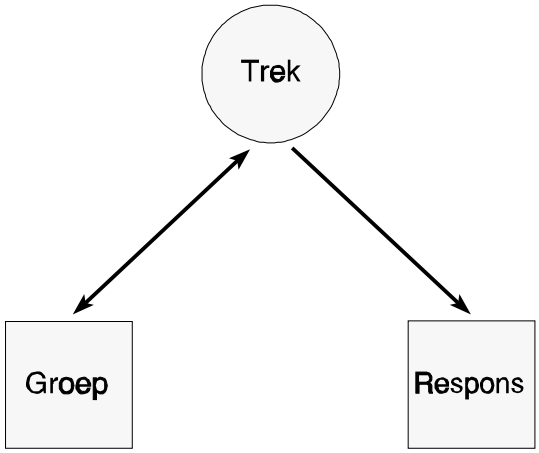
9.1 Definitie van onzuiverheid

Een algemene omschrijving van het begrip onzuiverheid, die zowel van toepassing is op het niveau van tests als van vragen, wordt gegeven door Mellenbergh (1985). In deze omschrijving wordt uitgegaan van een samenhang tussen groepslidmaatschap en de respons op een vraag of de score op een test. Men kan hierbij bijvoorbeeld denken aan het verband tussen het al dan niet behoren tot de groep autochtone leerlingen en de

score op een schooltoets. De relatie tussen groepslidmaatschap en de respons op een item of een toetsscore wordt in figuur 9.1 schematisch weergegeven, waarbij de geobserveerde variabelen (groepslidmaatschap en de respons) zijn aangegeven als blokken en de samenhang tussen die variabelen is aangeduid als een pijl met twee punten. Deze pijl geeft aan dat er sprake is van een samenhang tussen de variabelen en niet van een specifieke invloed van de ene variabele op de andere.



Figuur 9.1
Samenhang tussen groepslidmaatschap en respons



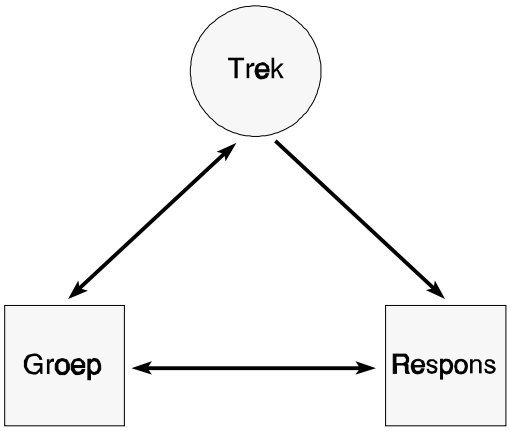
Figuur 9.2
Een zuivere vraag of test

Een dergelijke samenhang tussen groepslidmaatschap en respons hoeft niet te duiden op onzuiverheid van de vraag of de test, maar kan ook het gevolg zijn van werkelijke niveauverschillen tussen de betreffende groepen. Dit wordt weergegeven in figuur 9.2. Daar wordt de samenhang tussen het groepslidmaatschap en de respons geheel verklaard door een latente, niet direct geobserveerde variabele, een latente trek. De latente variabele is weergegeven als een cirkel en de invloed van deze variabele op de respons met een pijl met één punt. Omdat de verschillen op de vraag of de test veroorzaakt zijn door werkelijke vaardigheidsverschillen spreekt men van een zuivere vraag of test.

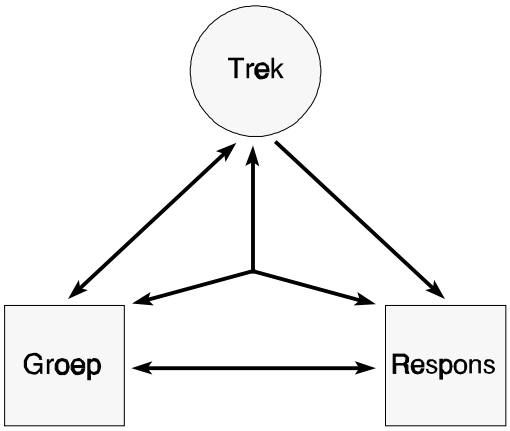
Er is sprake van een onzuivere vraag of test als de verschillen tussen de groepen niet helemaal verklaard kunnen worden door verschillen op de latente vaardigheidsdimensie. Dit wordt weergegeven in figuur 9.3, waar naast de samenhang tussen het groepslidmaatschap en de latente trek en de invloed van de latente trek op de respons nog steeds een directe samenhang blijft bestaan tussen het groepslidmaatschap en de

respons. Wanneer deze verschillen in prestaties tussen de groepen voor alle niveaus van de latente trek gelijk zijn, spreekt men van een uniform-onzuivere vraag of test.

Het is echter ook mogelijk dat de verschillen tussen de groepen variëren over de verschillende niveaus van de latente trek. Dit is bijvoorbeeld het geval als bij een laag vaardigheidsniveau de ene groep leerlingen hoger scoort terwijl bij een hoog vaardigheidsniveau de andere groep leerlingen hoger scoort. In deze situatie spreekt men van een niet-uniform onzuivere vraag. Niet-uniforme onzuiverheid wordt weergegeven in figuur 9.4, waarbij de drie pijlen vanuit het midden aangeven dat er sprake is van een samenhang tussen groepslidmaatschap en de respons welke gerelateerd is aan het niveau van de latente trek (een samenhang tussen de drie variabelen samen).



Figuur 9.3
Een uniform onzuivere vraag of test

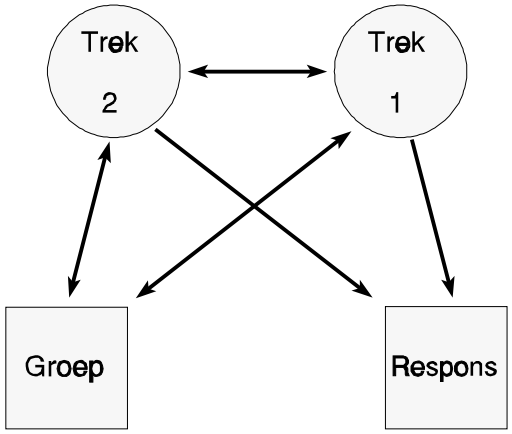


Figuur 9.4
Een niet-uniform onzuivere vraag of test

Tenslotte wordt in figuur 9.5 de situatie weergegeven waarbij de onzuiverheid verklaard wordt door het beschouwen van een tweede latente variabele, die niet tot de oorspronkelijke meetpretentie van het betreffende instrument hoort. Deze tweede latente variabele veroorzaakt de samenhang tussen het groepslidmaatschap en de respons. Na het toevoegen van deze trek is de samenhang tussen de geobserveerde variabelen, het groepslidmaatschap en de respons, verdwenen.

Wat betreft de hiervoor gegeven algemene beschrijving van het begrip onzuiverheid is het niet van belang of de geobserveerde respons op één of enkele vragen van een test, of op een hele test betrekking heeft. Bij het ontwikkelen van een methodologie voor het opsporen en aantonen van onzuiverheid is het daarentegen wel relevant of een test in zijn geheel onzuiver is, of dat slechts enkele vragen onzuiver zijn. Als een test

in z'n geheel onzuiver is, moet men om het groepseffect te kunnen evalueren namelijk over een additionele meting beschikken die wel zuiver is. Bij deze additionele meting moeten de groepsverschillen voldoende verklaard worden door verschillen op de latente trek. Wanneer de assumptie van normaliteit van de testscores aannemelijk kan worden gemaakt doordat bijvoorbeeld het scorebereik van de test voldoende groot is zodat de variabelen bij benadering continu zijn, kunnen variantie- of factoranalytische modellen worden toegepast. In het geval van één of enkele onzuivere vragen ligt het probleem anders, omdat daar naast de onzuivere ook zuivere vragen in de test aanwezig zijn. Aangezien de scores op testvragen echter meestal dichotoom of polytoom zijn, zal de assumptie van normaliteit per vraag meestal niet aannemelijk kunnen worden gemaakt. De itemresponstheorie levert in dat geval een meer geëigende context voor het ontwikkelen van een methodologie voor het opsporen en aantonen van onzuiverheid.



Figuur 9.5

Een onzuivere vraag of test waarbij onzuiverheid veroorzaakt wordt door één extra latente variabele

9.2 Methoden voor het bepalen van vraagonzuiverheid

In het onderzoek naar onzuiverheid is het gebruikelijk onderscheid te maken tussen een referentiegroep, zeg de meerderheidsgroep, en de potentieel benadeelde groep, die wordt aangeduid als de doelgroep. Wanneer bijvoorbeeld onzuiverheid als gevolg van culturele verschillen onderzocht wordt, bestaat de referentiegroep over het algemeen

uit autochtone en de doelgroep uit allochtone leerlingen. Deze terminologie zal ook in het vervolg van dit hoofdstuk gehanteerd worden.

Als we de theorie uit de vorige paragraaf vertalen naar dichotome items, is vraagonzuiverheid of DIF te definiëren als de omstandigheid dat bij een gegeven vaardigheidsniveau twee willekeurige leden van twee verschillende populaties niet dezelfde kans hebben om een vraag goed te maken. De statistische technieken voor het opsporen van DIF zijn dan ook alle gebaseerd op het evalueren van verschillen tussen de groepen in de kansen op een goed antwoord, conditioneel op een of andere maat voor vaardigheid. Meestal neemt men als maat voor de vaardigheid de somscore van de leerlingen. De meest algemeen toegepaste technieken zijn gebaseerd op de Mantel-Haenszel-toets (Holland & Thayer, 1988) of op IRT-modellen (Hambleton & Rogers, 1989; Kelderman, 1989). In de volgende twee paragrafen worden deze twee benaderingen toegelicht, in de daaropvolgende paragraaf worden zij met elkaar vergeleken. Daarna zal een concreet voorbeeld van het opsporen van vraagonzuiverheid met een itemresponsmodel worden gegeven.

9.2.1 De Mantel-Haenszel-procedure

Holland en Thayer (1988) stellen de volgende procedure voor om vast te stellen of de verschillen tussen de groepen in de moeilijkheidsgraad van een item, conditioneel op de somscores van de leerlingen, statistisch significant zijn. Voor elke niveaugroep, dat wil zeggen voor elke groep leerlingen met een score in een bepaald bereik, wordt een 2x2-tabel van itemscore bij groepslidmaatschap opgesteld. De tabel is weergegeven in figuur 9.6, waarbij in de cellen de aantallen personen staan aangegeven.

	Score op item i		
	1 (goed)	0 (fout)	Totaal
Referentiegroep	a_q	b_q	n_{1q}
Doelgroep	c_q	d_q	n_{2q}
Totaal	m_{1q}	m_{0q}	n_q

Figuur 9.6
2x2-tabel van niveaugroep q

Betekenis van de symbolen in figuur 9.6:

n_q totaal aantal kandidaten in niveaugroep q;

a_q personen in de referentiegroep bij niveaugroep q die item i juist beantwoord hebben;

b_q personen in de referentiegroep bij niveaugroep q die item i onjuist beantwoord hebben;

c_q personen in de doelgroep bij niveaugroep q die item i juist beantwoord hebben;

d_q personen in de doelgroep bij niveaugroep q die item i onjuist beantwoord hebben.

De door Holland en Thayer voorgestelde procedure is gebaseerd op een zogenaamde 'odds-ratio' (ratio van kansen) α_q. Deze wordt geschat door

$$\hat{\alpha}_q = \frac{p_{1q}/(1-p_{1q})}{p_{2q}/(1-p_{2q})} = \frac{a_q d_q}{b_q c_q}, \tag{9.1}$$

waarbij p_{1q} de kans op een goed antwoord is van de referentiegroep en p_{2q} de kans op een goed antwoord van de doelgroep. Wanneer de prestaties van beide groepen niet verschillen, is $\hat{\alpha}_q$ gelijk aan 1. In het geval de twee groepen verschillende antwoordpatronen vertonen, is $\hat{\alpha}_q$ groter dan 1 wanneer de referentiegroep een grotere kans op een goed antwoord heeft en $\hat{\alpha}_q$ kleiner dan 1 wanneer dit voor de doelgroep geldt. Voor de Mantel-Haenszel-toets worden de Mantel-Haenszel-statistieken van alle niveaugroepen gecombineerd tot

$$\hat{\alpha}_{MH} = \frac{\sum_q a_q d_q / n_q}{\sum_q b_q c_q / n_q}. \tag{9.2}$$

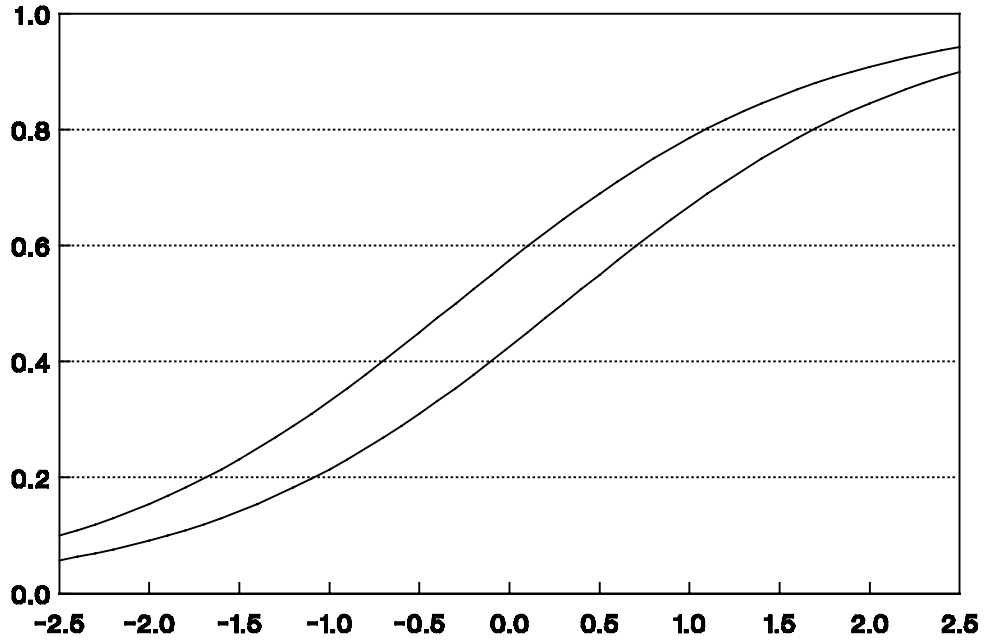
Indien er in de populaties geen DIF voorkomt en dus $\alpha_{MH} = 1$, kan aangetoond worden dat $\log\hat{\alpha}_{MH}$ normaal verdeeld is met een gemiddelde nul en standaarddeviatie $\mathrm{SE}(\log\hat{\alpha}_{MH})$, zodat de gestandaardiseerde log-odds-ratio $z = \log\hat{\alpha}_{MH}/\mathrm{SE}(\log\hat{\alpha}_{MH})$ bij benadering standaard-normaal is verdeeld. Bij een significantie-niveau van 1%, zijn de kritische waarden $z \geq 2.58$ als het item gemakkelijker is in de referentiepopulatie en $z \leq -2.58$ als het item moeilijker is in de referentiepopulatie.

De aanwezigheid van items met DIF doet afbreuk aan de waarde van de somscore als indicator van de vaardigheid van de leerlingen. De somscore wordt immers mede bepaald door items die voor de twee groepen een verschillende moeilijkheidsgraad hebben. Daarom is het zoeken naar DIF een iteratief proces. Eerst wordt een analyse

uitgevoerd waarbij de antwoorden op alle items worden opgenomen in de somscore. Vervolgens wordt er een analyse uitgevoerd waarbij de items die in de eerste analyse een significante uitkomst van de Mantel-Haenszel-toets hadden niet meer in de somscore worden opgenomen. Nu is het enerzijds mogelijk dat er nieuwe items met significante DIF bijkomen, anderzijds is het mogelijk dat de significante DIF verdwijnt bij items die in de eerste analyse wel een significante uitkomst van de Mantel-Haenszel opleverden. Het iteratieve proces gaat door tot er een verzameling items zonder DIF gevonden wordt waarmee de somscore berekend kan worden en een verzameling items met een significante uitkomst van de Mantel-Haenszel-toets die niet in de berekening van de somscore zijn betrokken.

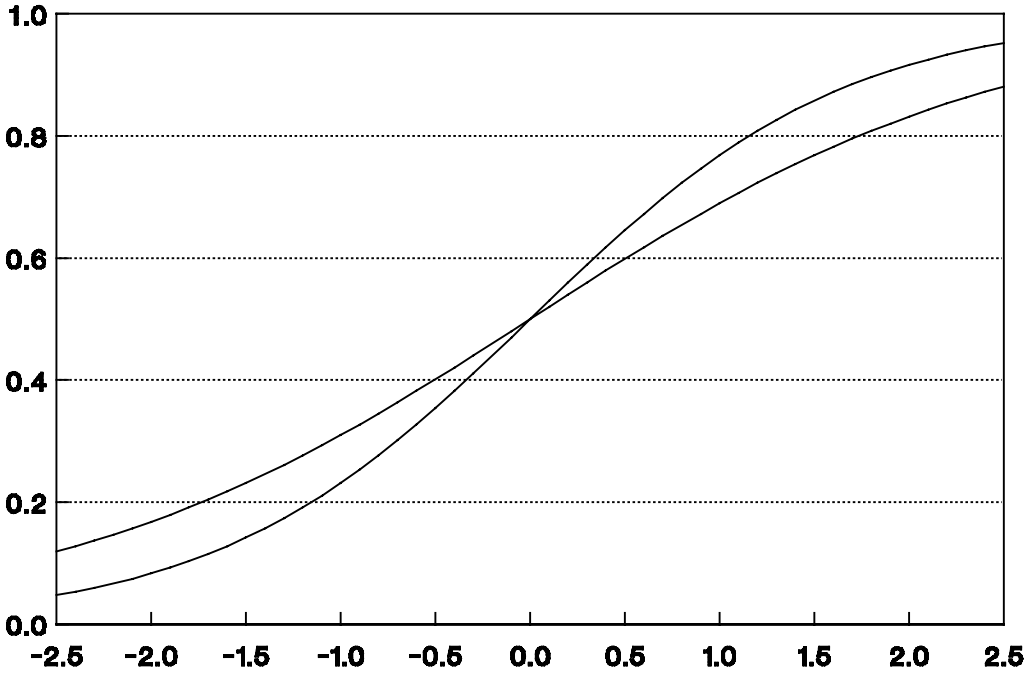
9.2.2 Procedure met IRT-modellen

In de itemresponstheorie wordt de kans op een goed antwoord op een item beschreven als een functie van persoonsparameters en itemparameters. Deze eigenschap maakt de klasse van IRT-modellen bijzonder geschikt voor het onderzoeken van DIF: conditioneren op het vaardigheidsniveau van respondenten is hier niets anders dan het constant houden van de persoonsparameters. Individuen met gelijke persoonsparameters moeten, ongeacht de populatie waartoe ze behoren, dezelfde kans op een goed antwoord hebben. Items kunnen verschillen in moeilijkheidsgraad en groepen kunnen verschillen in hun bekwaamheid om een juist antwoord op een item te geven, maar dat is op zich nog geen vraagonzuiverheid. Een item wordt alleen als onzuiver beschouwd als de moeilijkheidsgraad ervan varieert tussen personen van eenzelfde vaardigheidsniveau die tot verschillende populaties behoren. De generalisatie van DIF naar polytome items volgt eenvoudig uit de definitie voor dichotome items: een polytoom item is onzuiver als de verzameling van kansen om in één van de categorieën van het item te scoren, conditioneel op het vaardigheidsniveau, verschilt tussen groepen. Bij deze definities is niet van belang welk itemresponsmodel bij de data past. De term vaardigheidsniveau kan bijvoorbeeld betrekking hebben op een multidimensionale vaardigheidsparameter θ, zoals die voorkomt in het Raschmodel met een multivariate vaardigheidsverdeling dat behandeld is in hoofdstuk 5. Een unidimensionaal IRT-model maakt de problematiek conceptueel echter een stuk eenvoudiger.



Figuur 9.7
Responsfunctie van een uniform onzuiver dichotoom item

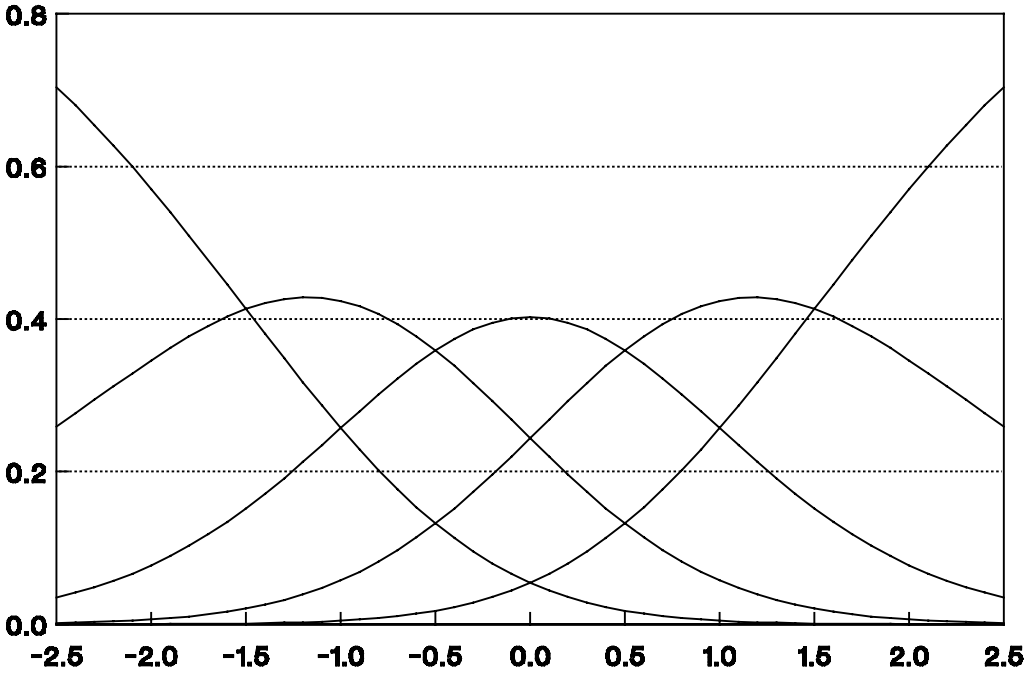
In paragraaf 9.1 is een onderscheid gemaakt tussen uniforme en niet-uniforme onzuiverheid. Een dichotoom item is uniform onzuiver als de kans op een juist antwoord in de doelpopulatie voor alle vaardigheidsniveaus lager is dan in de referentiepopulatie, of als het omgekeerde het geval is. Een voorbeeld van een dergelijk item wordt gegeven in figuur 9.7. Een item is niet-uniform onzuiver als de kans op een juist antwoord voor verschillende vaardigheidsniveaus nu eens in het voordeel is van de referentiepopulatie en dan weer in het voordeel is van de doelpopulatie. Een voorbeeld daarvan wordt gegeven in figuur 9.8.



Figuur 9.8
Responsfunctie van een niet-uniform onzuiver dichotoom item

In het voorbeeld doet de ene groep het op een laag vaardigheidsniveau beter dan de andere, terwijl dit op een hoog vaardigheidsniveau precies omgekeerd is. De systematische patronen van figuur 9.7 en 9.8 kunnen goed gemodelleerd worden door de locatie- en discriminatieparameters van het item te variëren over de groepen. In de praktijk kan het patroon van onzuiverheid veel onregelmatiger zijn en is het expliciet modelleren van de responsen van beide groepen niet altijd mogelijk.

De generalisatie van de concepten van uniforme- en niet-uniforme vraagonzuiverheid van dichotome naar polytome items is gecompliceerd omdat er in dat geval meer dan één itemresponsfunctie per item aanwezig is. In het voorbeeld van figuur 9.9 geeft de monotoon dalende curve links de kans op een score in de categorie nul aan, terwijl de monotoon stijgende curve rechts de kans op een score in de hoogste categorie aangeeft. De resterende eentoppige curven geven de kansen aan om in de overige categorieën te scoren. De itemresponscurven voldoen aan het partial credit model (PCM), maar aangezien slechts één item wordt beschouwd voldoen ze tevens aan het één-parameter logistische model (OPLM). In het PCM zijn de parameters β_{ij}, $j = 1,...,m_i$ de grenswaarden waar de kansen om in de categorie $j-1$ en de categorie j te scoren, gelijk zijn. Dat wil zeggen, de parameters geven de positie op de x-as aan waar de curven van categorie $j-1$ en j elkaar snijden.



Figuur 9.9

Itemresponsfunctie in het partial credit model

Het onderscheid tussen uniforme en niet-uniforme vraagonzuiverheid is intuïtief gezien bij dichotome items gerelateerd aan het al dan niet elkaar snijden van de itemkarakteristieke curven voor de verschillende populaties. In het geval van polytome items is een dergelijk eenvoudige definitie door het aantal karakteristieke curven en hun onderlinge afhankelijkheid niet mogelijk. Voor unidimensionale polytome modellen, zoals het PCM, het rating scale model of het OPLM kan men een item uniform onzuiver noemen wanneer de verwachte score op het item gegeven θ in de doelpopulatie systematisch hoger of lager is dan in de referentiepopulatie.

Onderzoek naar vraagonzuiverheid met behulp van IRT

Zoals hiervoor in termen van IRT is aangegeven, is een item onzuiver als de kansen op de responsen in de categorieën van het item, conditioneel op het vaardigheidsniveau, tussen groepen verschillen. De procedure voor het aantonen van dit verschijnsel bestaat uit twee stappen:

(1) het zoeken naar een passend IRT model voor de data van de referentiegroep en, voor zover mogelijk, de doelgroep,

(2) het evalueren van de verschillen in responskansen tussen de referentie- en de doelgroep in homogene subgroepen van gelijke vaardigheid.

Indien onzuivere items gevonden worden, kan men nog twee bijkomende stappen zetten: (3) het modelleren van de responsen van de doelpopulatie op de onzuivere items,

(4) het evalueren van de consequenties van de aanwezigheid van DIF, door het schatten van de resultaten (bijv. de scoreverdeling) van de doelpopulatie voor het geval geen DIF aanwezig zou zijn.

Met betrekking tot de eerste stap is allereerst de keuze van een itemresponsmodel van belang. Bij veel toetsen wordt de meting uitgevoerd door gebruik te maken van een ongewogen somscore. Dit betekent dat men de leerlingen ordent op een unidimensionaal vaardigheidscontinuüm en dat de persoonsparameter unidimensionaal is. Fischer (1974, pp. 193-203) heeft aangetoond dat onder de assumptie dat de somscore een voldoende steekproefgrootheid is voor een unidimensionale vaardigheidsparameter, en een paar technische assumpties (lokale stochastische onafhankelijkheid, een strikt monotoon stijgende kans op een goed antwoord die nergens gelijk aan nul of een is), het Raschmodel noodzakelijkerwijze volgt. Met andere woorden, het gebruik van de somscore als uitkomst van de met het toetsinstrument uitgevoerde meting impliceert dat de resultaten van de meting in feite aan het Raschmodel zouden moeten voldoen. Vaak voldoen de data echter niet aan het Raschmodel en moet men gebruik maken van andere modellen zoals het OPLM of een model met een multivariate vaardigheidsverdeling. Dit betekent dat de responskansen op de items conditioneel op de door deze modellen voorgeschreven steekproefgrootheden voor de vaardigheidsparameters moeten worden geëvalueerd. Met andere woorden, de rol van het IRT-model is het leveren van een adequate beschrijving van de vaardigheid van de leerlingen. In dit verband zullen we hier kort ingaan op een door Bügel en Glas (1991) gerapporteerd onderzoek naar vraagonzuiverheid bij examens tekstbegrip voortgezet onderwijs. Voor de eerste stap van het onderzoek, het zoeken naar een passend IRT-model voor de data van de referentiegroep en, voor zover mogelijk, de doelgroep, maakten zij gebruik van een variant van het model met een multivariate vaardigheidsverdeling dat beschreven is in hoofdstuk 5. Om zo dicht mogelijk bij de uiteindelijke resultaatbepaling van de examens te blijven, werd door de onderzoekers in de verzameling opgaven van het complete examen eerst gezocht naar een aantal Rasch-homogene subsets van items. Voor ieder van die subschalen is de somscore een voldoende grootheid voor de vaardigheidsparameter. In de examensituatie worden, voor de uiteindelijke resultaatbepaling, de somscores op de subschalen opgeteld tot een totaalscore als eindwaardering. Dit impliceert in feite een, meestal arbitraire, waardering voor de verschillende vaardigheidsdimensies: bij een andere combinatie van deelscores tot een

eindwaardering ontstaat namelijk een andere ordening van leerlingen. Overigens is de correlatie tussen de vaardigheidsdimensies hoog (altijd groter dan .85) zodat de afwijking ten opzichte van het Raschmodel niet bijzonder groot is en men zeker niet mag concluderen dat een examen een aantal scherp afgebakende vaardigheidsdimensies meet. Men zou de gevonden multidimensionaliteit eerder kunnen kenschetsen als additionele ruis bij een unidimensionaal Raschmodel. Het door Bügel en Glas gekozen IRT-model is niet per definitie het enig juiste. De essentie van de eerste stap is het zoeken van een passend IRT-model om een adequate maat voor de vaardigheid van de leerlingen te construeren. Zo zal voor het voorbeeld in dit hoofdstuk een andere keuze gemaakt worden, en zal gebruik worden gemaakt van het OPLM. Voor meer informatie over de procedure met het Raschmodel met een multivariate vaardigheidsverdeling zij men verder verwezen naar Bügel en Glas (1991).

De tweede stap van het onderzoek naar onzuiverheid is het evalueren van de verschillen in responskansen tussen de referentie- en doelgroep in subgroepen van gelijke vaardigheid. Hieronder zal worden beschreven hoe dit, in het kader van het OPLM, kan worden uitgevoerd. Hiertoe zullen twee toetsen voor het OPLM, de R_{1c}- en de S_i-toets, worden aangepast voor het opsporen van vraagonzuiverheid.

Om het zoeken van een passend IRT model niet te laten beïnvloeden door eventueel aanwezige onzuivere items, is het verstandig in eerste instantie alleen de gegevens van de referentiegroep te gebruiken. Voor het evalueren van de modelpassing kan men gebruik maken van de in de hoofdstukken 4 en 5 beschreven toetsen. Als een voor de referentiegroep passend model gevonden is, breidt men de analyse uit naar beide groepen. Stel dat groepslidmaatschap wordt aangeduid met het subscript g, waarbij de referentiegroep wordt geïndiceerd met $g = 1$ en de doelgroep met $g = 2$. Zoals bij de eerder geïntroduceerde versies van de R_{1c}- en S_i-toets (zie formule 5.44 en 5.45) worden homogene niveaugroepen, geïndexeerd met q, gevormd op basis van de voldoende statistieken s voor de persoonsparameters. Dus net als in de hoofdstukken 4 en 5 bestaat niveaugroep q uit alle leerlingen die een score s in een scorebereik G_q hebben. Beide toetsen zijn gebaseerd op het verschil tussen de proportie antwoorden in categorie j van item i in scoregroep s, $p_{ij|s}$ en de onder het model geschatte kans op een antwoord in categorie j van item i in scoregroep s, $\hat{\pi}_{ij|s}$. Voor het evalueren van vraagonzuiverheid worden deze proporties en kansen voor iedere groep g afzonderlijk uitgerekend, dus de toets zal nu gebaseerd zijn op proporties $p_{ij|sg}$ en geschatte kansen $\hat{\pi}_{ij|sg}$. De CML schattingen van de itemparameters worden berekend met behulp van de gegevens van zowel de referentie- als de doelgroep. Er wordt dus verondersteld dat voor beide groepen hetzelfde model geldt.

Om de relatie met de Mantel-Haenszel-procedure wat duidelijker te kunnen maken zullen we de veralgemening van de R_{1c}- en S_i-toets in termen van tellingen geven. Daartoe definiëren we de stochastische variabele $M_{ij|sg}$, met realisatie $m_{ij|sg}$, als het aantal antwoorden in categorie j van item i gegeven door personen van groep g en scoregroep s. De passing van het model voor beide groepen zal dus geëvalueerd worden met behulp van de verschillen tussen de geobserveerde en verwachte waarden van $M_{ij|sg}$. Deze verschillen zijn gegeven door

$$d_{ij|sg} = m_{ij|sg} - \mathcal{E}(M_{ij|sg} \mid \hat{\beta}) \tag{9.3}$$

waarbij $\mathcal{E}(M_{ij|sg} \mid \hat{\beta})$ de verwachte waarde is van $M_{ij|sg}$, uitgerekend met CML schattingen van de itemparameters β. Er geldt dat $m_{ij|sg} = n_{sg} p_{ij|sg}$ en $\mathcal{E}(M_{ij|sg} \mid \hat{\beta}) = n_{sg} \hat{\pi}_{ij|sg}$, met n_{sg} het aantal personen in groep g dat score s haalt. Naar analogie van (5.44) kan de globale modelpassing worden geëvalueerd met behulp van de asymptotisch chi-kwadraat verdeelde toetsingsgrootheid R_{1c}. Deze wordt benaderd door

$$R_{1c}^{*} = \sum_{g=1}^{2} \sum_{q=1}^{r} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\left[\sum_{s \in G_q} d_{ij|sg}\right]^2}{\sum_{s \in G_q} var(d_{ij|sg})}, \tag{9.4}$$

waarbij $var(d_{ij|sg})$ de variantie van het verschil $d_{ij|sg}$ is.

Merk op dat in het geval van dichotome items het aggregatieniveau van de data waarop de verschillen $d_{ij|sg}$, met $j = 1$, gebaseerd zijn, hetzelfde is als bij de Mantel-Haenszel-toets. Met de verschillen $d_{ij|sg}$ gaat men na of de proportie goede antwoorden voor de referentie- en doelgroep conform de voorspellingen van het model zijn en, omdat voor beide groepen hetzelfde model geldt, of deze proporties gelijk zijn. Als de toetsingsgrootheid significant is, is door inspectie van de verschillen $d_{ij|sg}$ na te gaan of de verwerping toe te schrijven is aan systematische verschillen tussen de twee groepen in de kans op het produceren van een goed antwoord. Per item kan men de verschillen $d_{ij|sg}$ ook combineren tot een toetsingsgrootheid die is op te vatten als een veralgemening van de itemgerichte S_{ij}-toets. De benaderende toetsingsgrootheid gedefinieerd door (5.45) wordt daartoe veralgemeniseerd tot

$$S_{ij}^{*} = \sum_{g=1}^{2} \sum_{q=1}^{r} \frac{\left[\sum_{s \in G_q} d_{ij|sg}\right]^2}{\sum_{s \in G_q} var(d_{ij|sg})}, \qquad (j = 1,...,m_i). \tag{9.5}$$

Als is aangetoond dat één of meer items in een toets onzuiver zijn, is de derde stap in het onderzoek naar DIF mogelijk. Deze stap heeft betrekking op de vraag of het antwoordgedrag van de doelgroep adequaat kan worden beschreven door een itemresponsmodel. Inzicht in de aard van de onzuiverheid is uiteraard essentieel voor het voorkomen ervan. Zowel bij dichotome als bij polytome items kan het variëren van locatie- en discriminatieparameters van het item soms voldoende zijn om het antwoordgedrag van de verschillende populaties te modelleren. Een voorbeeld hiervan wordt in paragraaf 9.3 gegeven. Er zijn echter uiteraard ook vormen van DIF denkbaar waarbij de onzuiverheid complexer van aard is. Zo is het bijvoorbeeld mogelijk dat onzuiverheid ten nadele van de doelgroep alleen bij lage vaardigheidsniveaus voorkomt, en dat bij hogere vaardigheidsniveaus de doelgroep zijn achterstand volledig weet te compenseren.

De vierde mogelijke stap in het onderzoek naar vraagonzuiverheid is het evalueren van de invloed van de onzuiverheid op de verdeling van zowel de gewogen als de ongewogen somscores van de respondenten. Daarvoor moet eerst de vaardigheidsverdeling van de referentiegroep en de vaardigheidsverdeling van de doelgroep geschat worden. Hiertoe kan men bijvoorbeeld het OPLM uitbreiden met de veronderstelling dat de vaardigheidsparameters in beide groepen, overigens verschillende, normale verdelingen hebben. Vervolgens kan men de parameters in dit uitgebreide model met behulp van MML schatten. Het is echter ook mogelijk de CML schattingen van de itemparameters als constanten te beschouwen en alleen ML-schattingen van de populatieparameters te maken. In beide gevallen is het echter wel noodzakelijk dat de passing van het uitgebreide model aannemelijk wordt gemaakt. De effecten van de aanwezigheid van DIF zijn nu als volgt te evalueren.

Stel dat N_{sg} het aantal respondenten van groep g is dat een gewogen of ongewogen score s haalt. Gegeven nu de schattingen $\hat{\beta}$ van de itemparameters en $\hat{\mu}_g$ en $\hat{\sigma}_g$, voor $g = 1$ en 2, van de populatieparameters, kan men voor alle mogelijke scores s de verwachte waarde $\mathscr{E}(N_{sg} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g)$ berekenen. Dit is overigens geen triviale aangelegenheid. Stel dat $\{\boldsymbol{x} \mid s\}$ de verzameling is van alle mogelijke antwoordpatronen $\boldsymbol{x}$ die resulteren in een score s. Dan berekent men deze verwachte waarden als

$$\mathscr{E}(N_{sg} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g) \;=\; N_g \sum_{\{\boldsymbol{x} \mid s\}} P(\boldsymbol{x} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g). \qquad (9.6)$$

Met andere woorden, men moet de kansen op antwoordpatronen sommeren over alle antwoordpatronen die resulteren in score s. Doordat ook hier echter symmetrische basisfuncties een rol blijken te spelen (zie Glas, 1991) is dit echter minder bewerkelijk

dan het lijkt. Het gaat er nu om, de resultaten van de doelpopulatie te schatten als de toets geen onzuivere items had gehad, met andere woorden, als de itemparameters voor beide groepen gelijk zouden zijn geweest. Daartoe kan men de verwachte frequentieverdeling van de doelpopulatie $\mathscr{E}(N_{sg} \mid \hat{\beta},\hat{\mu}_g,\hat{\sigma}_g)$ berekenen met voor de onzuivere items parameterwaarden die gevonden zijn bij de referentiepopulatie.

9.2.3 De relatie tussen de Mantel-Haenszel-procedure en de IRT-procedure

Een speciaal geval van de hierboven geschetste procedure met behulp van itemresponsmodellen is die welke gebaseerd is op het Raschmodel voor dichotome items. Zowel deze procedure als die met de Mantel-Haenszel-toets zijn allebei gebaseerd op hetzelfde principe, namelijk het toetsen of de kans op een goed antwoord gegeven een somscore of een bereik van somscores hetzelfde is voor de referentie- en de doelgroep. Beide technieken hebben voordelen en hun beperkingen.

Bij de Mantel-Haenszel-procedure is de somscore, in tegenstelling tot bij het Raschmodel, niet gevalideerd als maat voor de vaardigheid van de respondenten. Het gebruik van de ongewogen somscore is overigens niet essentieel voor de Mantel-Haenszelprocedure. Ook is het mogelijk om de niveaugroepen voor de toets op basis van een andere statistiek voor vaardigheid te vormen. Hierbij kan men bijvoorbeeld denken aan een gewogen somscore zoals bij OPLM gebruikt wordt. Ook hier blijft echter de kritiek dat deze maat voor het vaardigheidsniveau eerst gevalideerd zou moeten worden.

Een andere nadeel van de Mantel-Haenszel-procedure is dat niet alle vormen van onzuiverheid gedetecteerd kunnen worden. In het geval van uniforme onzuiverheid is de kans op een goed antwoord voor één van de groepen over het hele scorebereik systematisch hoger. In het geval van niet-uniforme onzuiverheid zijn er niveaus waarop de ene groep en niveaus waarop de andere groep beter scoort. De Mantel-Haenszelprocedure is alleen gevoelig voor de eerste vorm van onzuiverheid, in het tweede geval vallen de effecten in de toetsstatistiek tegen elkaar weg. De toetsingsgrootheden voor het Raschmodel en OPLM leiden niet aan dit euvel omdat hier de verschillen tussen verwachte en geobserveerde proporties gekwadrateerd worden.

Het toepassen van het Raschmodel of OPLM heeft echter als nadeel dat de parameterschatting leidt tot restricties op de toetsingsgrootheden, waardoor een item met DIF ten nadele van de ene groep kan resulteren in één of meer items die schijnbaar DIF vertonen ten nadele van de andere groep. Dit ongewenste effect ontstaat doordat de CML schattingsvergelijkingen voor de itemparameters te schrijven zijn als

$$\sum_g \sum_s m_{ij|sg} = \sum_g \sum_s \mathscr{E}(M_{ij|sg} \mid \beta), \tag{9.7}$$

zodat, na invulling van de schattingen geldt dat $\Sigma_{g,s}\ d_{ij|sg} = 0$. Met andere woorden, voor ieder item is de som over groepen respondenten van de verschillen tussen verwachte en geobserveerde frequenties nul. Dit betekent dat door de schattingsmethode, vraagonzuiverheid die de ene groep benadeelt altijd samengaat met een bevoordeling van de andere groep. Restrictie (9.7) geldt voor ieder item afzonderlijk. Er ontstaan door de schattingsmethode echter ook afhankelijkheden die betrekking hebben op alle items. Na CML schatting geldt namelijk ook dat $\Sigma_i\ d_{ij|sg} = 0$, met $j = 1$. Dus voor iedere groep respondenten is de som over items van de verschillen tussen verwachte en geobserveerde frequenties ook nul. Voor iedere groep respondenten wordt de aanwezigheid van benadelende items hierdoor vertaald in de aanwezigheid van bevoordelende items, vice versa.

Gezien deze overwegingen is het raadzaam de beide technieken zo veel mogelijk in elkaars verlengde te hanteren. Zo kan men bijvoorbeeld eerst Rasch-homogene subschalen of een passend OPLM zoeken en op de aanwezigheid van DIF toetsen met het IRT model, om vervolgens voor iedere subschaal de Mantel-Haenszel-techniek toe te passen. Door deze vorm van kruisvalidatie kan men artefacten die samenhangen met de gebruikte methode zoveel mogelijk vermijden.

9.2.4 Een voorbeeld van het bepalen van vraagonzuiverheid met behulp van OPLM

Het voorbeeld dat gegeven zal worden betreft een deel van het eindexamen HAVO voor het vak economie. Dit voorbeeld vormde een onderdeel van een groter onderzoek naar geslachtsgebonden vraagonzuiverheid bij de eindexamens in het voortgezet onderwijs. Aangezien het hier de bedoeling is om statistische procedures te illustreren en niet om inhoudelijk op de uitkomsten van het onderzoek naar vraagonzuiverheid in te gaan, zullen geen voorbeelden van onzuivere items getoond worden of conclusies worden getrokken over de mate waarin het verschijnsel voorkomt.

De analyses werden uitgevoerd op een steekproef van 1000 jongens en 1000 meisjes uit de totale examenpopulatie. Voor de eenvoud van de presentatie zal het voorbeeld tot tien polytoom gescoorde items beperkt worden.

De eerste stap van de procedure bestond uit het zoeken van een passend OPLM. Dit gebeurde door een iteratieve procedure van het postuleren van discriminatie-indices, het berekenen van CML schattingen, het toetsen en bijstellen van de hypothesen met betrekking tot de discriminatie-indices. Om het zoeken naar een geschikt model niet

te laten beïnvloeden door mogelijk aanwezige DIF, zijn eerst alleen de data van de referentiegroep gebruikt. De analyses werden uitgevoerd met het computerprogramma OPLM (Verhelst, Glas & Verstralen, 1993). In tabel 9.1 wordt een overzicht gegeven van de uitkomsten van de toetsen voor het definitieve model. In de kolom "A" worden de discriminatie-indices weergegeven.

Tabel 9.1

Overzicht van passingstoetsen voor de referentiegroep

Item	A		S	df	P	M	M2	M3
1	2	[:1]	11.724	7	.110	-.294	-.648	-.039
		[:2]	6.685	7	.462	-.460	.098	-.584
2	3	[:1]	5.918	6	.432	-1.390	.716	.587
		[:2]	6.346	7	.500	-.195	.554	.029
		[:3]	4.025	5	.546	.003	.512	.878
3	4	[:1]	9.685	5	.085	1.543	2.476	3.615
		[:2]	1.624	6	.951	.893	.750	.167
4	2	[:1]	4.054	7	.774	.578	.423	.163
		[:2]	10.543	7	.160	.238	-.309	-1.202
		[:3]	3.582	5	.611	.472	.010	-.634
5	2	[:1]	9.124	6	.167	1.408	1.601	1.888
		[:2]	2.208	7	.947	.284	.837	-.631
		[:3]	5.140	7	.643	-1.064	.494	-.928
6	3	[:1]	6.090	7	.529	.743	.761	.006
		[:2]	4.065	7	.772	.315	.836	.414
7	3	[:1]	5.873	7	.555	-.063	-.961	.286
		[:2]	15.456	6	.017	.528	-.645	1.892
8	3	[:1]	6.971	5	.223	-.687	-.361	-1.348
		[:2]	15.915	6	.014	-1.473	-.427	-2.709
		[:3]	6.283	6	.392	.010	-.002	-.141
9	4	[:1]	6.359	6	.384	.120	-.930	-.779
		[:2]	1.958	6	.923	-1.202	-.913	-.386
10	4	[:1]	2.321	4	.677	-.187	-1.186	-.158
		[:2]	2.575	5	.765	-1.126	-.794	-1.339
		[:3]	5.503	5	.358	-.653	-1.213	.532

R_{1c} = 75.182; df = 72; p = .3757

De splitsing van het scorebereik van een item in de scores $0,...,j$ en $j+1,...,m_i$ kan in verkorte notatie worden weergegeven als $[:j+1]$, voor $j = 0,...,m_i-1$. Het programma OPLM berekent de S_{ij}- en M-toetsen voor alle dichotomisaties [:1],...,$[:m_i]$. In de

kolom "S" worden de waarden van de S_{ij}-toetsen weergegeven, de volgende twee kolommen geven respectievelijk het aantal vrijheidsgraden en de overschrijdingskansen. In de laatste drie kolommen worden de waarden van de drie versies van de M-toets gegeven, deze toetsen zijn asymptotisch normaal verdeeld. Aan de hand van de waarde van de R_{1c}-toets die onderaan de tabel staat afgedrukt, kan men zien dat de passing van het model aanvaardbaar is. In de daarop volgende twee analyses werden de discriminatie-indices die voor de referentiegroep waren gevonden niet veranderd. In de eerste analyse werden CML schattingen berekend en modeltoetsingen uitgevoerd op de doelpopulatie. In de tweede analyse werden CML parameterschattingen en modelpassing berekend op beide groepen tegelijk. De resultaten van de daarbij behorende R_{1c}-toetsen staan vermeld in tabel 9.2 in de rijen genummerd twee en drie. Het blijkt dat het model in beide gevallen verworpen moest worden. De resultaten van de tweede analyse laten zien dat de discriminatie-indices van de referentiepopulatie niet passen in de doelpopulatie, zelfs wanneer de schattingen van de itemparameters in deze laatste groep verkregen zijn.

Tabel 9.2
Hypothesetoetsing

analyse	model	R_{1c}	df	prob
1.	referentiegroep	75.182	72	.3757
2.	doelgroep	127.283	72	.0001
3.	gecombineerde groepen	356.747	168	.0000
4.	doelgroep, 9 aangepaste index	59.982	72	.8430
5.	gecombineerde groepen, 3 gesplitst	258.614	166	.0000
6.	gecombineerde groepen, 9 gesplitst	379.550	166	.0000
7.	gecombineerde groepen, 3 en 9 gesplitst . .	154.301	164	.6971

De resultaten van de derde analyse geven ook aan dat de gecombineerde data van beide groepen tegelijk, niet goed door hetzelfde model beschreven kunnen worden. Om na te gaan of dit laatste resultaat een gevolg is van DIF wordt in tabel 9.3 een overzicht gegeven van de passingstoetsen voor beide groepen samen. De tabel heeft hetzelfde formaat als tabel 9.1. Het blijkt dat de items drie en negen in belangrijke mate bijdragen aan het niet passen van het model. Onderaan de tabel staat de bijdrage van de twee groepen aan de uitkomst van de R_{1c}-toets. De bijdrage van de doelgroep (een χ^2 van 212.64) is veel groter dan de bijdrage van de referentiegroep (een χ^2 van 144.11).

Gezien het feit dat de discriminatie-indices bepaald zijn op de referentiegroep is dit niet verwonderlijk.

Om de hypothese van DIF verder te onderzoeken, kunnen bijvoorbeeld de verschillen tussen geobserveerde en verwachte frequenties behorend bij de R_{1c}-toets geïnspecteerd worden. Voor het berekenen van deze toets zijn de respondenten van zowel de referentie- als van de doelgroep, op basis van hun gewogen somscores, opgedeeld in vier subgroepen. Deze subgroepen werden zodanig samengesteld dat ze ongeveer hetzelfde aantal respondenten bevatten. De gekozen scoreniveaus en de resulterende aantallen respondenten per subgroep staan vermeld in de eerste twee regels van tabel 9.4. Verder worden voor alle items en alle categorieën de gestandaardiseerde afwijkingen tussen de verwachte en de geobserveerde frequenties in de subgroepen getoond. Voor de interpretatie van deze getallen is het belangrijk in gedachte te houden dat het realisaties van bij benadering standaard normaal verdeelde variabelen zijn.

Tabel 9.3

Overzicht van passingstoetsen voor de doel- en referentiegroep samen

Item	A		S	df	P	M	M2	M3
1	2	[:1]	28.189	14	.013	-.864	-.791	-1.121
		[:2]	12.748	14	.546	.067	.517	-1.236
2	3	[:1]	7.399	11	.766	-.070	.183	1.079
		[:2]	13.011	14	.526	-.838	-.625	-1.210
		[:3]	4.268	10	.934	.795	.755	.024
3	4	[:1]	107.862	12	.000	2.315	.658	2.771
		[:2]	37.500	12	.000	-1.438	.548	-1.787
4	2	[:1]	8.121	14	.883	-.721	-.351	-.338
		[:2]	15.971	14	.315	-.131	-.475	-.610
		[:3]	15.665	10	.110	-.137	-1.084	-1.317
5	2	[:1]	11.393	12	.496	1.428	-.339	.395
		[:2]	15.399	14	.351	-1.318	-1.453	-1.701
		[:3]	10.520	14	.723	-1.997	-1.455	-1.384
6	3	[:1]	10.486	14	.726	.358	1.505	.543
		[:2]	11.375	14	.656	.442	1.264	.518
7	3	[:1]	18.279	14	.194	-.438	-1.395	-.066
		[:2]	18.005	12	.116	1.376	-1.221	1.179
8	3	[:1]	9.410	10	.494	-1.049	-.234	-.955
		[:2]	19.127	13	.119	-1.341	-.615	-1.566
		[:3]	8.080	12	.779	-.322	.173	-.609
9	4	[:1]	113.760	12	.000	4.025	4.297	4.614
		[:2]	35.874	12	.000	2.657	2.655	3.173
10	4	[:1]	14.893	9	.094	-1.120	-1.083	-1.070
		[:2]	16.264	10	.092	-1.642	-1.612	-2.343
		[:3]	24.262	11	.012	-2.164	-2.123	-.712

groep	#items	#subgr.	#deviaties	R_{1c}

1	10	4	96	144.11
2	10	4	96	212.64

R_{1c} = 356.747; df = 168; p = .0000

Aan het teken kan men zien of er meer of minder observaties waren dan voorspeld door het model. In de kolommen "SS" worden de kwadratensommen van de afwijkingen vermeld, voor alle combinaties van items en categorieën. Merk op dat met name de kwadratensommen van item 3 groot zijn vergeleken met de kwadratensommen van de andere items. Verder vallen de geschaalde afwijkingen voor de referentiegroep over het algemeen positief uit, terwijl de afwijkingen bij dit item voor de doelgroep negatief zijn.

Tabel 9.4

Geschaalde afwijkingen op grond van CML schattingen, verkregen in beide groepen tegelijk

		referentiegroep					doelgroep				
		1	2	3	4		1	2	3	4	
Range →		1-20	21-37	38-52	53-73		1-20	21-36	37-52	53-73	
#obs →		228	237	253	263	SS	240	246	239	252	SS
Item	cat										
1	1	-.1	-1.5	.6	-.6	3.2	1.3	-.9	.3	1.1	4.0
	2	-1.5	.4	-.9	-.1	3.5	.5	.2	.7	.3	1.0
2	1	-.9	-.1	.8	.4	1.7	1.9	-.2	-.2	-2.0	7.8
	2	-.4	-.5	-1.2	-.0	2.1	-.2	-.0	.8	1.3	2.5
	3	-1.4	.9	.4	-.8	3.8	-.1	.3	-.0	.2	.1
3	1	5.1	2.8	.0	-1.0	35.7	-2.9	-3.4	.5	-1.0	21.6
	2	.9	2.4	1.7	.4	10.4	-1.8	-2.0	-3.2	1.0	18.9
4	1	1.9	-.0	.5	.0	4.1	-.4	-.6	-1.7	.6	4.0
	2	-2.0	-.8	.1	-.3	4.8	-.2	2.2	1.8	-1.3	10.3
	3	-.8	-.2	-1.0	-.0	1.8	-.8	-.2	-.8	2.0	5.6
5	1	1.8	.4	1.0	-1.1	6.0	-.8	-.1	-.8	-1.4	3.4
	2	.0	-.1	-.8	-.0	.7	-.6	.8	-.1	1.1	2.4
	3	-.8	-1.1	-.2	.5	2.4	.7	-.8	1.4	.5	3.4
6	1	-.3	.6	.4	.7	1.2	-.4	.2	-.2	-1.3	2.1
	2	-.4	-.4	-1.3	-.3	2.3	.4	1.7	-.1	.7	3.7
7	1	-.7	.0	-1.0	1.1	2.8	-.9	.6	1.0	-.5	2.6
	2	1.5	-2.8	.6	-1.1	11.8	-.2	.8	.8	.6	1.9
8	1	1.8	-.5	.5	-1.1	5.2	.2	-.9	-1.0	.6	2.4
	2	-2.3	1.0	.4	-.0	6.6	-.1	.2	.6	-.7	1.0
	3	-.7	.5	-1.3	.7	3.2	1.2	-.4	.0	.2	1.7
9	1	1.1	.4	.6	-1.0	3.0	1.4	-.2	-1.9	.1	6.0
	2	.7	.4	1.0	1.7	4.8	3.3	.7	-1.0	-4.2	30.6
10	1	.4	1.1	.6	-.6	2.4	-.2	-.4	-1.6	.0	3.0
	2	-1.3	-.0	.2	.8	2.6	-1.5	1.9	-.4	-1.5	8.6
	3	-1.0	-2.0	-.5	-.4	5.8	.6	-.0	1.9	1.5	6.6
SS →		63.0	36.8	18.3	15.0	133.1	39.2	34.4	36.8	45.8	156.3

$R_{1c}\rightarrow$	58.5	45.2	20.4	19.9	144.1	39.2	45.3	43.8	84.2	212.6

Dat wil zeggen dat dit item de referentiegroep bevoordeelt, aangezien deze groep meer responsen in de categorieën $h>0$ vertoont dan op grond van een in beide groepen samen gecalibreerd model verwacht zou kunnen worden. Op dezelfde wijze is het item nadelig voor de doelgroep, aangezien deze groep minder responsen in de categorieën $j > 0$ vertoont, en dus meer responsen in categorie $j = 0$. Voor item 9 is het patroon veel minder duidelijk.

Op grond van de analyse die in tabel 9.2 met een 2 genummerd is, zou verwacht kunnen worden dat de discriminatie-index voor item 9 in beide groepen verschillend zou zijn. Daartoe werd de analyse uitgevoerd die in tabel 9.2 met een 4 genummerd is. Voor deze analyse, waarbij alleen de gegevens van de doelgroep gebruikt weren, werd de discriminatie-index voor dit item van 4 in 2 veranderd. In tabel 9.2 is te zien dat deze aanpassing inderdaad resulteerde in een goede modelpassing: de uitkomst van R_{1c} is 59.982 bij 72 vrijheidsgraden.

In de laatste drie analyses waarvan de resultaten van de hypothesetoetsing in tabel 9.2 vermeld staan, is getracht om een model te construeren wat voor de data van beide groepen tegelijk zou passen. In analyse 5 is toegelaten dat de parameters van item 3 voor de referentie- en de doelgroep verschillend zouden kunnen zijn, waarbij de discriminatieparameter constant is gehouden. Dit resulteerde echter niet in een acceptabele modelpassing. In analyse 6 werd dezelfde procedure toegepast voor item 9, met dit verschil dat de discriminatie-index in de referentiegroep op vier werd gezet en in de doelgroep op twee. Opnieuw waren de resultaten onbevredigend. Tenslotte werd in analyse 7 voor beide items toegelaten dat de moeilijkheidsparameter tussen de groepen zouden kunnen verschillen en dit bleek, zoals te zien in de laatste regel van tabel 9.2, in een acceptabele modelpassing te resulteren. Resumerend kan men stellen dat item 3 uniform onzuiver is, omdat de itemparameters per groep verschillen, terwijl de discriminatie per groep gelijk is, terwijl item 9 niet-uniform onzuiver is, omdat ook de discriminatie-index aangepast moest worden. Overigens werden item 3 en 9 ook in de Mantel-Haenszel-procedure als onzuiver geïdentificeerd. Hiermee is de derde stap in het onderzoek, het modelleren van de responsen van de doelpopulatie afgesloten.

Tot slot werd de vierde stap van het onderzoek naar vraagonzuiverheid gezet door het evalueren van de invloed van de onzuiverheid op de verdeling van zowel de gewogen als de ongewogen somscores van de respondenten. Als eerste stap werd daartoe de passing van het model uit analyse 7, uitgebreid met normale vaardigheidsverdelingen voor de referentie- en doelgroep, onderzocht. De

itemparameters β en populatieparameters μ_g en σ_g, voor $g = 1$ en 2, werden geschat met behulp van MML. Berekening van de R_0-toets (zie hoofdstuk 4) resulteerde in een waarde van 121.79 (df: 138, p: .83), terwijl het berekenen van R_{1m} een waarde 267.82 opleverde (df: 303, p: .92), zodat dit uitgebreide model niet verworpen hoefde te worden. Hierna werd voor de doelpopulatie de frequentieverdeling $\mathscr{E}(N_{sg} \mid \hat{\beta}, \hat{\mu}_g, \hat{\sigma}_g)$ berekend met de parameters van de items 3 en 9 gelijk aan de waarden die gevonden werden bij de referentiepopulatie en de schattingen van de populatieparameters van de doelpopulatie. Op deze wijze worden de resultaten van de doelpopulatie op een zuivere toets geschat, dat wil zeggen, de resultaten voor het geval de itemparameters voor de referentie- en doelpopulatie gelijk zouden zijn geweest. Deze geschatte frequentieverdeling op een zuivere toets kan men dan vervolgens vergelijken met de gerealiseerde frequentieverdeling. Voor het bovenstaande voorbeeld werden de berekeningen uitgevoerd voor zowel de gewogen als de ongewogen scores. In beide gevallen bleek het gemiddelde van de verwachte frequentieverdeling voor de doelpopulatie lager voor de onzuivere test. Het verschil bedroeg overigens in beide gevallen minder dan één scorepunt. Met andere woorden de onzuiverheid had inderdaad een bescheiden negatieve invloed op het gemiddelde resultaat van de doelpopulatie.

9.3 Conclusie

Itemresponstheorie biedt een goed gefundeerd kader voor het opsporen van vraagonzuiverheid. Hierbij is het echter belangrijk dat de hulpmiddelen die de IRT ons aanreikt ook zorgvuldig worden gebruikt. In de eerste plaats dient een passend IRT-model te worden gevonden. Hierbij spelen twee aspecten een rol: de data en de mate waarin de passing van de verschillende IRT-modellen statistisch goed gefundeerd te evalueren zijn. Het OPLM beschikt enerzijds over een goed uitgerust toetsingsarsenaal en blijkt anderzijds in veel gevallen goed bij de data te passen. Daar komt bij dat de statistische toetsen voor dit model zo zijn te generaliseren, dat ze gevoelig zijn voor vraagonzuiverheid. Door parameterschatting en andere oorzaken kan de informatie die de toetsen opleveren enigszins vertroebelen. Daarom is het aan te bevelen de resultaten te kruisvalideren door het uitvoeren van een Mantel-Haenszel-procedure, waarbij de niveaugroepen gevormd worden op basis van de afdoende statistieken van het passende IRT-model. Tenslotte is een niet onaantrekkelijk aspect van het werken met een IRT-model dat men het niet hoeft te laten bij het opsporen van vraagonzuiverheid, maar dat men ook de effecten hiervan op de toetsresultaten kan schatten.