# 5

# Een overzicht van itemresponsmodellen

In hoofdstuk 4 is uitvoerig ingegaan op het Raschmodel, waarbij de nadruk vooral kwam te liggen op de statistische aspecten van schatting en toetsing. Het is niet zo dat dit model een allesoverheersende plaats inneemt in de IRT-literatuur. Er zijn zeer veel IRT-modellen ontwikkeld, en een volledig overzicht geven van de bestaande modellen is in het bestek van een hoofdstuk niet mogelijk. De selectie die zal worden gepresenteerd, weerspiegelt naast een zekere voorkeur van de auteurs, enkele aspecten die voor de praktijk belangrijk zijn, eerder dan diepe theoretische overwegingen. Een van de aspecten is het algemeen beschikbaar zijn van computerprogrammatuur.

Thematisch valt dit hoofdstuk uiteen in twee onderdelen, die men zou kunnen aanduiden als specificatie en generalisatie van het Raschmodel. Nadere specificatie van het Raschmodel is het antwoord op de vraag 'wat kun je verder nog doen als het Raschmodel bij de data past' en generalisatie van het Raschmodel is het antwoord op de vraag 'wat te doen indien het Raschmodel niet bij de gegevens past ?'.

Indien het Raschmodel overtuigend bij de data past, hoeft dit niet noodzakelijkerwijze het einde van de psychometrische bemoeienissen met deze data te betekenen. Naast de praktische toepassingsmogelijkheden van een deugdelijke schaal, kan men zich ook de vraag stellen hoe het komt dat het ene item moeilijker is dan het andere. Dit wil zeggen dat men probeert een theorie te construeren die de verschillen in moeilijkheid tussen de items verklaart. Binnen de IRT is een benadering ontworpen die toelaat een grote klasse van deze theorieën formeel te beschrijven en statistisch te toetsen. Hoewel deze benadering in principe op elk IRT-model kan worden toegepast, heeft ze haar eerste en ook omvangrijkste uitwerking gekregen in het kader van het Raschmodel. Technisch gezien komt deze benadering neer op het opleggen van een aantal restricties aan de itemparameters. In hoofdstuk 4 is dit ook al een keer gedaan om de rationale van de LR- en de Wald-toetsen te beschrijven. Het resulterende model is minder algemeen dan het Raschmodel, en kan dus worden opgevat als een nadere specificatie ervan. Deze specificatie weerspiegelt een bepaalde theorie of hypothese

over de structuur van de moeilijkheid van de items. Een gedetailleerde uiteenzetting van dit model is het onderwerp van paragraaf 5.1

Indien het Raschmodel niet bij de data past, kan men twee standpunten innemen. Men kan items of personen verwijderen totdat de overblijvende items zich wel adequaat door het Raschmodel laten beschrijven. Daarbij kan echter de inhoudsvaliditeit van de toets of de generaliseerbaarheid naar de populatie van personen in het gedrang komen. Men kan ook proberen te achterhalen waarom het model niet past. In hoofdstuk 4 hebben we gezien dat het Raschmodel gelijke discriminatie van de items veronderstelt. Als we erachter komen, bijvoor- beeld met behulp van de $M_i$ – toetsen, dat niet-passing toe te schrijven is aan ongelijke discri- minatie, kunnen we het Raschmodel vervangen door een algemener model dat ongelijke discriminatie toelaat, zoals het tweeparameter logistisch model dat in hoofdstuk 4 reeds kort werd besproken.

Generalisatie van het Raschmodel heeft ook nog een andere motivatie. Indien men over items beschikt met antwoordvariabelen die niet twee maar drie of meer verschillende waarden aannemen, dan komt de variant van het Raschmodel uit het vorige hoofdstuk niet in aanmerking, zodat men wel gedwongen is van een ander model gebruik te maken. Terzijde dient opgemerkt te worden dat het bespreken van IRT-modellen als generalisaties van het Raschmodel als didactisch hulpmiddel wordt gehanteerd en niet overeenkomt met de feitelijke historische ontwikkeling van de IRT: veel van de te presenteren modellen zijn eerder ontwikkeld dan het eigenlijke Raschmodel.

Paragraaf 5.2 is gewijd aan een algemene bespreking van de indelingsprincipes van IRT-modellen. In de paragrafen 5.3 en 5.4 komen unidimensionale modellen voor respectievelijk dichotome en polytome items aan de orde. Paragraaf 5.5 bespreekt enkele multidimensionale modellen.

## 5.1 Het lineair-logistische testmodel

Veronderstel dat de items van een toets bestaan uit wiskundige functies waarvan de afgeleide functie gevraagd wordt. Voor het nemen van afgeleiden bestaan specifieke regels, zoals:

$$\frac{d\,x^{n}}{d\,x} = n\,x^{n-1}$$

en

$$\frac{d \ln x}{d x} = \frac{1}{x}.$$

Nu is de hypothese dat de moeilijkheid van de items afhangt van de moeilijkheid van deze regels. Fischer (1973) stelde een zeer eenvoudig model voor om aan te geven hoe de itemmoeilijkheid tot stand komt. Indien in item *i* regel 1 tweemaal moet worden toegepast en regel 2 driemaal, dan is de moeilijkheid van dit item gegeven door

$$\beta_i = 2\eta_1 + 3\eta_2,$$

waarin $\eta_1$ en $\eta_2$ de moeilijkheden van de twee regels voorstellen. De coëfficiënten 2 en 3 in de gelijkheid hierboven zijn bekende constanten die volgen uit een analyse van de items. Indien we nu een toets maken met $k > 2$ items, die allemaal alleen een beroep doen op deze twee regels, dan moeten niet *k* parameters geschat worden, maar slechts 2, omdat de *k* itemparameters allemaal lineaire functies zijn van de twee η-parameters. Deze η-parameters worden aangeduid als basisparameters of elementaire parameters.

De veralgemening van bovenstaand voorbeeld is erg eenvoudig. Indien er $d < k$ basisparameters zijn, is het model gegeven door

$$\beta_i = \sum_{j=1}^{d} q_{ij} \eta_j, \qquad (i = 1,...,k). \tag{5.1}$$

De coëfficiënten $q_{ij}$ in (5.1) zijn constanten die a priori in het model worden ingebracht en niet uit de data worden geschat. Deze coëfficiënten of gewichten zoals ze vaak worden genoemd, representeren dus de theorie van de onderzoeker. Formule (5.1) zegt dat de itemparameters lineaire combinaties zijn van *d* elementaire parameters en een dergelijke modellering wordt aangeduid als het lineair-logistische testmodel (LLTM). Dit model werd voorgesteld door Fischer (1974, 1983).

Het LLTM heeft dus twee componenten: de antwoorden op de items kunnen beschreven worden door het Raschmodel, en bovendien zijn de itemparameters specifieke lineaire combinaties van meer basale parameters η. Het schattingsprobleem zal dus bestaan uit het schatten van deze η-parameters en bij de toetsing moet de geldigheid van beide componenten van het model onderzocht worden. Schatting en toetsing worden hierna besproken.

### *5.1.1 Parameterschatting in het LLTM*

We beginnen met een onderzoek van de aannemelijkheidsfunctie. In het Raschmodel is de aannemelijkheidsfunctie gegeven door formule (4.28), die we hier herhalen:

$$\ln L(\beta,\theta;\boldsymbol{X}) = \sum_{v=1}^{n} s_v \theta_v + \sum_{i=1}^{k} t_i(-\beta_i) - \sum_{v=1}^{n}\sum_{i=1}^{k} \ln\left[1 + \exp(\theta_v - \beta_i)\right], \tag{5.2}$$

waarin

$$s_v = \sum_{i=1}^{k} x_{vi}, \qquad t_i = \sum_{v=1}^{n} x_{vi}.$$

Substitueren we nu het rechterlid van (5.1) voor $\beta_i$ in het rechterlid van (5.2), dan krijgen we:

$$\ln L(\eta,\theta;\boldsymbol{X}) = \sum_{v=1}^{n} s_v\theta_v + \sum_{j=1}^{d}(-\eta_j)\sum_{i=1}^{k} t_i q_{ij} - \sum_{v=1}^{n}\sum_{i=1}^{k} \ln\left[1 + \exp\left(\theta_v - \sum_{j=1}^{d} q_{ij}\eta_j\right)\right], \tag{5.3}$$

waarin we duidelijk de structuur van de exponentiële familie herkennen. De laatste term in het rechterlid is uitsluitend een functie van de parameters, de eerste term is onveranderd gebleven in vergelijking met (5.2), en de middelste term is een som van *d* produkten, waarvan een factor de parameter $\eta_j$ is. De andere factor, $\sum_i t_i q_{ij}$, is alleen een functie van de data. Deze factor is dus een voldoende steekproefgrootheid voor de parameter $\eta_j$, en het model behoort tot de exponentiële familie. Dit is trouwens een voorbeeld van een algemeen resultaat: indien een model behoort tot de exponentiële familie, dan behoort het speciale geval van dit model dat ontstaat door lineaire restricties op de parameters aan te brengen eveneens tot de exponentiële familie.

In (5.3) is bovendien, net als in het gewone Raschmodel, de somscore de voldoende steekproefgrootheid voor de persoonsparameter. Door te conditioneren op de score kunnen we de conditionele aannemelijkheidsfunctie opstellen. Omdat het LLTM een speciaal geval is van het Raschmodel, moet de algemene formule voor de conditionele aannemelijk- heidsfunctie die in hoofdstuk 4 werd gegeven, hier ook geldig zijn. De logaritme van deze aannemelijkheidsfunctie is gegeven door formule (4.43) die we hier herhalen:

$$\ln L(\eta;\boldsymbol{x}\,|\,\boldsymbol{s}) = \sum_{i} t_i \ln \varepsilon_i - \sum_{v} \ln \gamma_{s_v}(\boldsymbol{\varepsilon}) \tag{5.4}$$

waarin

$$\varepsilon_i = \exp(-\beta_i) = \exp\left(-\sum_j^d q_{ij}\eta_j\right). \tag{5.5}$$

Substitueren we nu het rechterlid van (5.5) in (5.4), dan krijgen we:

$$\ln L(\eta; \boldsymbol{x} \mid \boldsymbol{s}) = \sum_j^d (-\eta_j) \sum_i t_i q_{ij} - \sum_v \ln \gamma_{s_v}(\boldsymbol{\varepsilon}). \tag{5.6}$$

De schattingsvergelijkingen kunnen we opstellen door van (5.6) de partiële afgeleiden naar de η-parameters gelijk te stellen aan 0, maar we kunnen ook gebruik maken van een eigen- schap van de exponentiële familie, die inhoud dat de schattingsvergelijkingen gegeven zijn door de voldoende steekproefgrootheden gelijk te stellen aan hun verwachte waarde. Dan krijgen we als schattingsvergelijkingen:

$$\begin{aligned} \sum_i q_{ij} t_i &= \mathscr{E}\left[\sum_i q_{ij} T_i \mid s_v\right] \\ &= \sum_i q_{ij} \sum_v \mathscr{E}(X_{vi} \mid s_v) \\ &= \sum_i q_{ij} \sum_v \pi_{i \mid s_v}, \qquad (j = 1, \ldots, d). \end{aligned} \tag{5.7}$$

Een vergelijking met de CML-schattingsvergelijkingen (4.45) laat meteen zien dat het gewone Raschmodel ook beschouwd kan worden als een LLTM, door de coëfficiënten $q_{ij}$ te definiëren als

$$q_{ij} = \begin{cases} 1 \text{ indien } j = i \text{ en } i > 1, \\ 0 \text{ in andere gevallen}. \end{cases}$$

In het algemeen geldt dat in het LLTM de voldoende steekproefgrootheden gegeven zijn door $d$ lineaire combinaties van de itemtotalen $t_i$ en de schattingsvergelijkingen door het gelijk-stellen van die $d$ lineaire combinaties aan hun verwachte waarde. In het gewone Raschmodel geldt natuurlijk dat $d = k - 1$.

Eén probleem dient nog even aan de orde gesteld te worden, namelijk het probleem van de normering van de basisparameters. Bij de behandeling van het Raschmodel in hoofdstuk 4 hebben we gezien dat een van de itemparameters vrij kan worden gekozen, of iets algemener uitgedrukt, dat bij elke itemparameter een willekeurige constante $c$ kan worden opgeteld. Het LLTM is echter ook een Raschmodel en dus moet die vrijheid ook hier gelden. Dit is inderdaad zo, want de algemene vorm van het LLTM is iets algemener dan door (5.1) is aangegeven en luidt eigenlijk

$$\beta_i = \sum_{j=1}^{d} q_{ij}\eta_j + c, \qquad (i = 1,...,k), \tag{5.8}$$

waarin *c* ogenschijnlijk de status heeft van een parameter, maar niets anders is dan een willekeurige normalisatieconstante. In de afleidingen hierboven is gewerkt met (5.1) in plaats van met (5.8), doch dit is hetzelfde als de keuze $c = 0$; dat wil zeggen dat in alle afleidingen deze normering reeds was ingevoerd.

### *5.1.2 Het toetsen van het LLTM*

Bij het toetsen van het LLTM moeten we er rekening mee houden dat het model twee componenten heeft en dat het meestal zinvol is die twee componenten afzonderlijk te toetsen. Het heeft namelijk niet veel zin de geldigheid van de restricties (5.1) te toetsen, als het Raschmodel zonder die restricties niet houdbaar is. De eerste stap in de toetsing zal er dus uit bestaan dat het Raschmodel zonder restricties getoetst wordt. Dit impliceert dat de parameters in het algemene model geschat worden, waarna een of meer toetsen die in hoofdstuk 4 besproken zijn worden toegepast. Indien deze toetsen geen aanleiding geven het algemene model te verwerpen, kunnen we het Raschmodel zonder restricties gebruiken om een LR-toets te construeren. De vector met parameters in het algemene model is gegeven door $\varphi_u = (\beta_1,...,\beta_k)$ en in het beperkte model door $\varphi_r = (\eta_1,...,\eta_d)$. De toetsings- grootheid

$$2[\ln L^*(\varphi_u; X) - \ln L^*(\varphi_r; X)],$$

waarin $L^*$ het maximum van de conditionele aannemelijkheidsfunctie aanduidt, is asymp- totisch chi-kwadraat verdeeld met $k - 1 - d$ vrijheidsgraden. Details over de constructie van een LR-toets kan men vinden in paragraaf 4.3.3. Merk op dat (5.1) de nulhypothese is. Grote waarden van de toetsingsgrootheid geven dus aan dat de beperking van het model met de specifieke waarden $q_{ij}$ die gebruikt zijn, niet ondersteund wordt door de observaties. De coëfficiënten $q_{ij}$ maken dus deel uit van de nulhypothese en de reden tot verwerping van de nulhypothese zou dus kunnen zijn dat een of meer van die coëfficiënten verkeerd gespecificeerd zijn. We zullen hier een toets bespreken die gevoelig is voor zo'n verkeerde specificatie. In hoofdstuk 4 hebben we gezien dat om een LR-toets te construeren de parameters geschat moeten worden zowel in het algemene model als in het beperkte model. Bij de Wald-toetsen hoefden we maar één keer te schatten, namelijk onder het algemene model. De Wald-toetsen zijn gebaseerd op de rationele dat de restricties op de parameters in het beperkte

model ongeveer moeten gelden voor de parameterschattingen in het algemene model. Er bestaat echter ook een manier van toetsen waarbij de schatting van de parameters gebeurt onder het beperkte model. Deze toetsen staan in de literatuur bekend als Lagrange-Multiplier-toetsen (LM, Aitchison & Silvey, 1958) of efficiënte-score-toetsen (Rao, 1948). We geven hier een voorbeeld dat van toepassing is op het LLTM.

Stel dat we betwijfelen of we de coëfficiënt $q_{12}$ wel goed gespecificeerd hebben. Als we niet echt een uitgesproken idee hebben welke waarde die coëfficiënt moet aannemen, zouden we zijn waarde uit de data kunnen schatten. Maar dat betekent dat we het getal $q_{12}$ willen beschouwen als de waarde die een parameter, zeg $\kappa_{12}$, aanneemt. We veronderstellen dus een model dat als parameters niet alleen de $d$ $\eta$-parameters bevat, maar ook nog de extra parameter $\kappa_{12}$. We beschouwen dit model als het algemene model en de bijbehorende parametervector is gegeven door $\varphi_u = (\eta_1,...,\eta_d,\kappa_{12})$. Het beperkte model waaronder we de schatting hebben uitgevoerd, is een restrictie op de parameterruimte, want we hebben de parameter $\kappa_{12}$ gelijkgesteld aan de waarde $q_{12}$. Dus kunnen we schrijven: $\varphi_r = (\eta_1,...,\eta_d,q_{12})$. Het zal duidelijk zijn dat we voor een LR-toets of een Wald-toets met nulhypothese: $\kappa_{12} = q_{12}$ de parametervector $\varphi_u$ moeten schatten en dat is geen eenvoudige aangelegenheid. We weten dat de CML-schatter $\hat{\kappa}_{12}$ moet voldoen aan

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \kappa_{12}}\right|_{\kappa_{12} = \hat{\kappa}_{12}} = 0. \tag{5.9}$$

Deze betekent dat de partiële afgeleide, geëvalueerd op het punt van de CML-schatting, gelijk moet zijn aan nul. Indien nu de nulhypothese waar is, mag de schatting $\hat{\kappa}_{12}$ niet ver afwijken van de hypothetische waarde $q_{12}$ en moet dus gelden

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \kappa_{12}}\right|_{\kappa_{12} = q_{12}} \approx 0. \tag{5.10}$$

We hoeven dus de CML-schatting van $\kappa_{12}$ niet te berekenen, we moeten alleen de partiële afgeleide van de log-aannemelijkheidsfunctie evalueren op het punt $\kappa_{12} = q_{12}$. Die partiële afgeleide zal echter ook een functie zijn van de $\eta$-parameters en de waarden die we voor die parameters moeten invullen is in (5.10) niet aangegeven. De waarden die men voor de $\eta$-parameters invult, zijn hun CML-schattingen $\hat{\eta}_j$, $j = 1,...,d$, onder het beperkte model. De schattingen van alle $d + 1$ parameters

onder het beperkte model kunnen we dus aangeven als $\hat{\varphi}_r = (\hat{\eta}_1,...,\hat{\eta}_d, q_{12})$. Als de nulhypothese waar is moet dus ook gelden dat

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \kappa_{12}}\right|_{\varphi_u = \hat{\varphi}_r} \approx 0. \tag{5.11}$$

Merk op dat per definitie geldt dat

$$\left.\frac{\partial \ln L(\varphi_u; X \mid \boldsymbol{s})}{\partial \eta_j}\right|_{\varphi_u = \hat{\varphi}_r} = 0, \quad (j = 1,...,d). \tag{5.12}$$

Als we alle partiële afgeleiden van de log-aannemelijkheidsfunctie, geëvalueerd in het punt $\hat{\varphi}_r$ verzamelen in een $d + 1$ vector $\boldsymbol{b}(\hat{\varphi}_r)$, dan zijn de eerste $d$ elementen van die vector per definitie gelijk aan 0.

Stel dat we ook de matrix van tweede partiële afgeleiden naar alle $d + 1$ parameters van de vector $\varphi_u$ bepalen en evalueren in de waarden van $\hat{\varphi}_r$. Keren we het algebraïsche teken van deze matrix om, dan krijgen we de geobserveerde informatiematrix, geëvalueerd in $\hat{\varphi}_r$. Deze matrix kunnen we dus aanduiden als $I(\hat{\varphi}_r)$. De toetsingsgrootheid $LM(q_{12})$ is dan gegeven door

$$LM(q_{12}) = \boldsymbol{b}'(\hat{\varphi}_r)\,[I(\hat{\varphi}_r)]^{-1}\,\boldsymbol{b}(\hat{\varphi}_r) \tag{5.13}$$

en is onder de nulhypothese asymptotisch chi-kwadraat verdeeld met 1 vrijheidsgraad. Het uitrekenen van (5.13) is relatief eenvoudig omdat de elementen van de $\boldsymbol{b}$-vector die overeen- komen met de $\eta$-parameters exact gelijk zijn aan nul. Op deze vereenvoudiging gaan we hier echter niet in.

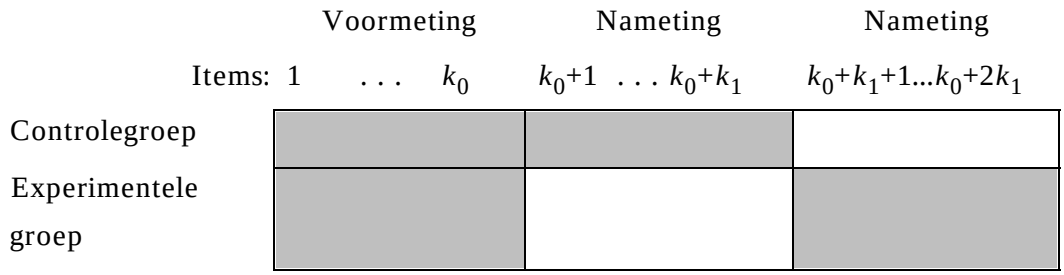
De LM-toetsen kunnen ook veralgemeend worden voor meer parameters tegelijkertijd, door in de $\boldsymbol{b}$-vector en in de informatiematrix de partiële afgeleiden op te nemen naar meerdere coëfficiënten $q_{ij}$ die men in de toetsing van de hypothese wil betrekken.

Hoewel het gebruik van de LM-toetsen zeer aantrekkelijk is voor verfijning van het LLTM, dienen toch een kanttekening gemaakt te worden. Deze kanttekening heeft te maken met een nuancering die we impliciet in de nulhypothese hebben ingebracht. De rationale van de LM-toets hebben we beschreven alsof het hele probleem eruit bestond te weten of de restrictie $\kappa_{12} = q_{12}$ waar was en daarbij hebben we gedaan alsof het

algemene model waar was. Maar dat algemene model is heel complex, het veronderstelt het Raschmodel en de lineaire restricties waarvan de coëfficiënten, met uitzondering van $q_{12}$, allemaal vaste waarden hebben. Deze gespecificeerde waarden maken dus ook deel uit van het algemene model en van het beperkte model. Indien een of meer van deze gespecificeerde waarden erg afwijken van de werkelijke waarden, is het onbeperkte model niet meer juist en is de toetsingsgrootheid $LM(q_{12})$ ook niet meer chi-kwadraat verdeeld. De LM-toetsen zijn dus vooral nuttig indien de restricties die aangebracht zijn niet al te ver bezijden de werkelijkheid zijn.

### *5.1.3 Een toepassing van het LLTM*

Een interessante toepassing van het introduceren van lineaire restricties op de itemparameters is het analyseren van gegevens die verzameld zijn in een experiment of een quasi-experiment. Stel dat in een experiment twee groepen worden onderscheiden: een experimentele groep die een behandeling krijgt en een controlegroep die geen behandeling krijgt. In beide groepen vindt een voor- en een nameting plaats. De voormeting wordt uitgevoerd met een toets van $k_0$ dichotome items en de nameting met een toets van $k_1$ items. De items in de voor- en de nameting behoeven niet dezelfde te zijn. Het is het meest voor de hand liggend om het effect van de behandeling te modelleren als een verandering in de persoonsparameters. Daar we echter gebruik willen maken van de in hoofdstuk 4 beschreven methodologische voordelen van de CML-schattingsmethode, zal in deze toepassing een verandering in de persoons- parameters vertaald worden in een verandering in de itemparameters. Met andere woorden, toename van de persoonsparameters wordt vertaald in een afname van de itemparameters. Als we aannemen dat de experimentele behandeling een positief effect heeft op de latente vaardigheid, moeten de itemparameters in de experimentele groep een kleinere waarde hebben dan in de controlegroep. Een elegante manier om dit te onderzoeken bestaat uit de volgende procedure, die logisch gezien twee stappen bevat. De eerste stap bestaat er uit, te doen alsof de oorspronkelijke $k_1$ items die voor de nameting worden gebruikt verdubbeld zijn, zodat er $2k_1$ items gebruikt zijn voor de nameting. Dit resulteert in een onvolledig design dat schematisch is weergegeven in figuur 5.1. De rijen in deze figuur zijn geassocieerd met groepen personen. De kolommen in de figuur zijn geassocieerd met items. Bij de voormeting hebben beide groepen dezelfde items gekregen. In de nameting is dat ook gebeurd, alleen hier wordt voorlopig even verondersteld dat de items door de experimentele manipulatie niet meer voor beide groepen hetzelfde zijn.




Figuur 5.1
Datamatrix met conceptuele items

Met andere woorden, elk 'fysiek' item in de nameting wordt gesplitst in twee 'conceptuele' items. We gaan er van uit dat de conceptuele items zo geordend zijn dat de conceptuele items $k_0 + i$ en $k_0 + k_1 + i$ naar hetzelfde fysieke item verwijzen. Deze associatie en de effecten van de behandeling worden nu gemodelleerd door het invoeren van de volgende lineaire restricties op de parameters van de conceptuele items:

$$\begin{cases} \beta_i = \eta_i, \quad (i = 1,...,k_0), \\ \beta_{k_0+i} = \eta_{k_0+i}, \quad (i = 1,...,k_1), \\ \beta_{k_0+k_1+i} = \eta_{k_0+i} + \tau, \quad (i = 1,...,k_1). \end{cases} \tag{5.14}$$

De associatie tussen de conceptuele items in de nameting komt tot uiting in de tweede en derde regel van (5.14) waar de twee conceptuele items $k_0 + i$ en $k_0 + k_1 + i$ betrokken worden op dezelfde basisparameter $\eta_{k_0+i}$. De parameter $\tau$ is de basisparameter die het effect van de experimentele behandeling weerspiegelt. Als $\tau$ positief is, worden de items moeilijker en heeft de experimentele behandeling dus een negatief effect. Bij een positief effect hoort een negatieve $\tau$. Het algebraïsche teken van $\tau$ wordt in (5.14) niet gespecificeerd. Om duidelijk te maken dat (5.14) een speciaal geval is van (5.1), kunnen we (5.1) herschrijven als een matrixvergelijking door alle $q_{ij}$'s op te vatten als de elementen van een $k \times d$ gewichtenmatrix $Q$:

$$\beta = Q\eta. \tag{5.15}$$

Passen we (5.15) nu toe op het bovenstaande voorbeeld met $k_0 = k_1 = 2$, dan krijgen we

$$\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \beta_5\\ \beta_6\end{bmatrix} = \begin{bmatrix}1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&1&0&1\\ 0&0&0&1&1\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\\ \tau\end{bmatrix}. \qquad (5.16)$$

Omdat we één itemparameter vrij kunnen kiezen, kunnen we bijvoorbeeld $\beta_1$ gelijkstellen aan 0, maar omdat $\beta_1 = \eta_1$, geldt dan dat $\eta_1 = 0$. Er zijn dus niet vijf vrije basisparameters maar slechts vier. De lineaire restricties op de vrije itemparameters krijgen we dus door in (5.16) de elementen $\beta_1$ en $\eta_1$ en de eerste rij van de matrix te schrappen.

Dit model kan getoetst worden door het opstellen van een LR-toets waarbij het algemene model de geldigheid van het Raschmodel voor alle $k_0 + 2k_1$ conceptuele items veronderstelt en waarbij dus $k_0 + 2k_1 - 1$ vrije β-parameters geschat worden. In het beperkte model, waar geschat wordt onder de restricties (5.14) zijn er $k_0 + k_1$ vrije basisparameters. De LR-toets levert dus een toetsingsgrootheid op die asymptotisch chi-kwadraat verdeeld is met $k_1 - 1$ vrijheidsgraden.

Als het model geldig is, betekent dit natuurlijk niet automatisch dat het experiment effect heeft gehad. Om dit aan te tonen moeten we de nulhypothese $\tau = 0$ toetsen. Dit kan door een Wald-toets te gebruiken, waarbij de toetsingsgrootheid gegeven is door $\hat{\tau}/SE(\hat{\tau})$ en die onder de nulhypothese asymptotisch standaardnormaal verdeeld is. Het toetsen van deze nulhypothese heeft alleen zin indien het gehanteerde LLTM houdbaar blijkt. Indien dit niet het geval is, heeft een toetsing van de effectparameter geen zin.

Bij de interpretatie van de resultaten moet uiteraard rekening worden gehouden met alle aspecten van de interne validiteit in het wetenschappelijk onderzoek; het gebruik van een IRT-model maakt methodologische overwegingen niet overbodig. Voor dit soort overwegingen zij men verwezen naar Campbell en Stanley (1966), we gaan er hier nu niet verder op in.

Indien de LR-toets een significant resultaat oplevert, zou men kunnen denken dat het gehanteerde LLTM te streng is en dat het wellicht versoepeld kan worden door niet één enkele τ-parameter in het model toe te laten, maar een, mogelijk verschillende, $\tau_i$-parameter voor elk item. Deze aanpak leidt echter tot logische problemen die verband houden met de proefopzet. Men gaat er namelijk van uit dat de hele verzameling gebruikte items aan het Raschmodel voldoen. Het Raschmodel schrijft echter voor dat de verandering in vaardigheid equivalent is met een en dezelfde verandering in de waarde van alle opgaven. Als men bij aparte items aparte effecten definieert, is het

bijvoorbeeld heel goed mogelijk dat de rangorde van de items op het latente continuüm voor de controle en de experimentele groep niet meer dezelfde is. Dit leidt dus tot een tegenspraak met de stelling dat alle opgaven aan het Raschmodel voldoen.

Tot slot van deze paragraaf nog een opmerking over de schatbaarheid van de parameters. Indien de voortoets weggelaten zou worden uit het design dat in figuur 5.1 is afgebeeld, zijn de parameters van het model, zowel met als zonder de restrictie (5.14) niet meer schatbaar. Men zou kunnen opperen dat dit rechtstreeks voortvloeit uit het in paragraaf 4.4 besproken feit dat CML-schattingen niet kunnen worden berekend uit een niet-verbonden design. Het probleem is in het algemeen echter iets gecompliceerder dan in paragraaf 4.4 werd besproken, omdat we het design moeten beschouwen in samenhang met de lineaire restricties. Zo kunnen er designs bestaan die zonder lineaire restricties niet schatbaar zijn, maar het wel worden met bepaalde lineaire restricties. De precieze condities wanneer dit het geval is, zijn gegeven in Fischer (1983). De conclusie is dus dat de voortoets niet kan worden weggelaten.

## 5.2 Indelingsprincipes van IRT-modellen

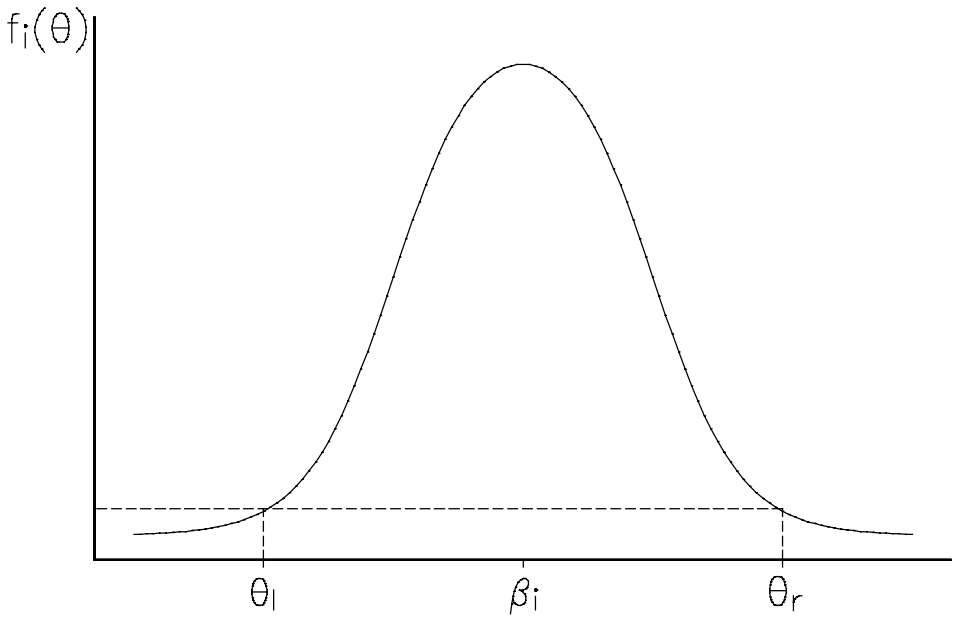
Om een inzicht te krijgen in de grote collectie IRT-modellen, zullen we drie indelingsprincipes hanteren: de algemene vorm van de itemresponsfunctie, namelijk monotoon tegenover niet-monotoon, het aantal categorieën dat de antwoordvariabele kan aannemen, namelijk twee tegenover meer dan twee, ofwel dichotoom tegenover polytoom en als derde de dimensionaliteit van de latente variabele. We becommentariëren kort deze drie principes.

In hoofdstuk 4 hebben we betoogd dat het een wenselijke eigenschap is van een IRT-model dat de itemresponsfunctie monotoon stijgend is in $\theta$: hoe groter de vaardigheid, des te groter de kans op een juist antwoord. We kunnen echter ook modellen beschouwen waarbij de latente variabele die we wensen te meten niet adequaat aangeduid wordt met de categorie ’vaardigheid’. Beschouw het volgende item uit een fictieve vragenlijst naar politieke interesse:

"Vindt U dat Joop den Uyl een goede premier van Nederland was ?",

waarbij een positief antwoord gecodeerd wordt met 1 en een negatief antwoord met 0. Indien we veronderstellen dat het antwoord op dit item bepaald wordt door de positie van de persoon op een continuüm dat de politieke ’links-rechts’-dimensie weerspiegelt, is het niet aannemelijk dat hoe rechtser de persoon is, hoe groter de kans zal zijn dat het item bevestigend beantwoord wordt. Een veel plausibeler model is grafisch

weergegeven in figuur 5.2, waarbij $\beta_i$ de positie van het item op het latente continuüm aangeeft.




Figuur 5.2
Een ééntoppige itemresponsfunctie

Deze positie weerspiegelt precies die politieke overtuiging die nodig is om de bovenstaande uitspraak met maximale kans te ondersteunen. De persoon met latente positie $\theta_l$ bevindt zich links van $\beta_i$ en heeft een kleine kans om het item bevestigend te beantwoorden: Den Uyl wordt te rechts bevonden. Een persoon met positie $\theta_r$ ja zegt met een even kleine kans, maar de reden is dat Den Uyl te links bevonden wordt. Modellen met een eentoppige in plaats van een monotone itemresponsfunctie horen thuis in een domein dat doorgaans wordt aangeduid met ontvouwingstheorie. Een uiteenzetting van deze theorie kan men vinden in het werk van haar grondlegger C.H. Coombs (1964). Een goed overzicht van verschillende IRT-modellen met eentoppige itemresponsfuncties vindt men in het aan ontvouwing gewijde themanummer van het tijdschrift Kwantitatieve Methoden (Hoijtink, 1993). Deze modellen komen in dit hoofdstuk verder niet meer ter sprake.

Bij de modellen met monotone itemresponsfuncties kan men een belangrijke onderverdeling maken volgens het soort wiskundige functie dat men hanteert. In het Raschmodel is dat bijvoorbeeld de logistische functie. De grafiek van deze functie lijkt echter erg op de grafiek van de (cumulatieve) normale verdelingsfunctie. Deze laatste functie is dan ook in veel modellen gebruikt. Deze modellen staan bekend onder de algemene naam 'normaal-ogiefmodellen'. Voor een algemene inleiding en een

rechtvaardiging van het gebruik van de normale-verdelingsfunctie, verwijzen we naar hoofdstuk 16 van Lord en Novick (1968). Hoewel de logistische functie bij wiskundige afleidingen tot veel eleganter resultaten leidt dan de normale verdelingsfunctie, wordt die laatste nog steeds gebruikt, zij het niet zozeer in de literatuur die men gewoonlijk onder de benaming IRT aanduidt, maar meer in het onderzoeksdomein van de structurele modellen; zie bijvoorbeeld Muthén (1984, 1987).

Een zeer opmerkelijke klasse van modellen ontstaat indien men probeert de specifieke vorm van de itemresponsfunctie zo weinig mogelijk vast te leggen. Bij de modellen met een logistische functie of bij het normaal-ogiefmodel wordt de familie van de itemresponsfuncties zodanig gespecificeerd dat alleen nog één of meer parameters moeten worden geschat om de functies volledig te kennen. Mokken (1971) heeft een klasse van modellen gespecificeerd waarbij alleen zeer algemene kenmerken van de itemresponsfuncties worden vastgelegd, zoals monotoniciteit en dat de grafieken van de functies elkaar niet snijden. Parameters komen daarbij niet voor en deze modellen worden dan ook vaak aangeduid als niet-parametrische IRT-modellen. Mokken heeft aangetoond dat met dit soort zwakke eisen toch zinvolle uitspraken over de θ-waarde van personen kunnen worden gedaan en dat eveneens statistisch kan getoetst worden of aan deze eisen wel voldaan is. Recent onderzoek naar niet-parametrische IRT-modellen kan men vinden in Sijtsma en Molenaar (1987). Van de modellen die verder in dit hoofdstuk worden besproken, behoren de itemresponsfuncties allemaal tot de familie van de logistische functies.

Het tweede indelingsprincipe heeft betrekking op het aantal antwoordcategorieën. Indien dit aantal groter dan twee is, spreekt men niet van dichotome items maar van polytome items. Het is belangrijk op te merken dat het kenmerk dichotoom versus polytoom te maken heeft met het aantal waarden dat de antwoordvariabele $X_i$ kan aannemen en dat dit aantal niet hetzelfde hoeft te zijn als het aantal categorieën waarin de oorspronkelijke observaties zijn ingedeeld. Een goed voorbeeld van dit onderscheid is het geval van meerkeuze-items. Stel dat een item met vier antwoordalternatieven, A, B, C en D, heeft, waarbij B het juiste antwoord is. Als we ervan uitgaan dat iedere persoon precies één van die alternatieven kiest, zijn er dus vier mogelijke antwoorden op dit item. Maar daaruit volgt niet dat we de antwoorden op dit soort items moeten analyseren met een model voor polytome items. We kunnen immers de oorspronkelijke observaties reduceren tot dichotome data door een punt toe te kennen indien het juiste alternatief gekozen is en geen punten in de andere drie gevallen. Indien we de versie van het Raschmodel uit hoofdstuk 4 gebruiken, analyseren we dichotome data en de statistische toetsen hebben alleen op deze data betrekking. Indien het model goed bij de data past, volgt daar niet uit dat deze analyse van de dichotome de enig juiste is.

Het is bijvoorbeeld mogelijk dat het kiezen van alternatief A een indicatie is van een grotere vaardigheid dan het kiezen van C of D. Indien we dit vermoeden hebben, kunnen we een analyse uitvoeren die gevoelig is voor dit onderscheid door een IRT-model voor polytome items te gebruiken. De wijze waarop de antwoorden van de personen gescoord worden, weerspiegelt een vermoeden of een hypothese en het gebruik van een formeel IRT-model is te beschouwen als een toetsing van deze hypothese. De geldigheid van een IRT-model betreft dus niet alleen de antwoorden (het gedrag) van de personen die de toets gemaakt hebben, maar ook de scoringsregel. De scoringsregel weerspiegelt een hypothese over de interpretatie die aan de responsen in de verschillende categorieën gegeven moet worden. In het boven- staande voorbeeld zouden we bijvoorbeeld 2 punten kunnen toekennen voor het antwoord B, 1 punt voor het antwoord A en 0 punten voor de antwoorden C en D, om vervolgens een model toe te passen waarbij een hogere itemscore als een indicator van een grotere vaardigheid wordt beschouwd. In dat geval zegt men dat we te doen hebben met een polytoom item met geordende antwoordcategorieën. Anderzijds zouden we ook de antwoorden A tot en met D ook kunnen omcoderen willekeurige getallen waarvan we de waarden niet wensen te interpreteren als geordende maar als nominale categorieën. Voor beide gevallen, geordende en nominale categorieën, zijn unidimensionele IRT-modellen ontwikkeld. Ze zullen behandeld worden in paragraaf 5.4.

Vooraleer we het derde indelingsprincipe bespreken, moeten we even ingaan op een complicatie die ontstaat wanneer de twee voorgaande indelingsprincipes met elkaar gecombineerd worden. Bij de bespreking van het eerste indelingsprincipe, monotone versus niet-monotone itemresponsfuncties, hebben we een terminologie gehanteerd die geschikt is voor dichotome items, maar die tekortschiet voor polytome items. Zoals we verder gedetail- leerd zullen bespreken, maar nu reeds intuïtief kunnen inzien, kunnen we voor een model met polytome items niet volstaan met een enkele itemresponsfunctie per item. We zullen een responsfunctie nodig hebben voor elke categorie van de antwoordvariabele. Daarom zullen we in het geval van polytome items ook niet meer spreken over de itemresponsfunctie maar over categorieresponsfuncties. Bovendien zal blijken dat niet alle categorieresponsfuncties van een item *i* monotoon stijgend of dalend in θ kunnen zijn. Om toch een indeling monotoon versus niet-monotoon te kunnen handhaven, zullen we de eigenschap monotoniciteit verder niet meer associëren met een categorieresponsfunctie, maar met een speciale functie die de itemregressiefunctie genoemd wordt. De regressie van de antwoordvariabele $X_i$ op de latente variabele θ is de verwachte waarde van $X_i$, beschouwd als een functie van θ. In het Raschmodel is die itemregressiefunctie gegeven door:

$$\mathcal{E}(X_i \mid \theta) = 1 \times f_i(\theta) + 0 \times [1 - f_i(\theta)] = f_i(\theta). \quad (5.17)$$

Bij dichotome antwoordvariabelen valt de itemregressiefunctie samen met de itemresponsfunctie. Bij polytome items kan de itemregressiefunctie beschouwd worden als een samenvatting van alle categorieresponsfuncties. We zullen van een monotoon item spreken indien de itemregressiefunctie van de antwoordvariabele monotoon is in θ, of, iets informeler uitgedrukt, het item is monotoon als een grotere vaardigheid een grotere verwachte itemscore impliceert.

Het derde indelingsprincipe is de dimensionaliteit van de latente variabele θ. In hoofdstuk 4 is er op gewezen dat de aanname van unidimensionaliteit centraal staat in het Raschmodel. Deze aanname betekent dat alle items in een toets dezelfde vaardigheid meten. Nu is het mogelijk dat de items in een toets een beroep doen op twee verschillende vaardigheden, maar niet allemaal in dezelfde mate. Anders gezegd, alle items doen een beroep op beide vaardigheden, maar de mate waarin kan voor beide vaardigheden van item tot item verschillen. Het is bijvoorbeeld aannemelijk dat redactiesommen in een rekentoets zowel een verbale als een numerieke vaardigheid aanspreken. Als ze dat in ongelijke mate doen, zal een unidimensionaal model waarschijnlijk niet toereikend zijn om het antwoordgedrag op een dergelijke toets adequaat te beschrijven. Men kan dan proberen de oorspronkelijke toets op te splitsen in twee unidimensionale deeltoetsen, bijvoorbeeld met behulp van Martin-Löfs toets voor unidimensionaliteit (zie paragraaf 4.3.1), of men kan een model gebruiken waarin de vaardigheid meerdimensionaal is.

Op het eerste gezicht lijkt een unidimensionaal model, zoals het Raschmodel, het allereenvoudigste geval in de klasse van multidimensionale modellen. Maar het concept van een enkele dimensie betekent dat verschillende θ-waarden zinvol kunnen worden geordend. Men kan deze ordening echter ook beschouwen als een te strenge eis en proberen een model te maken waarin de verschillende θ-waarden niet geordend zijn, maar worden behandeld als nominale categorieën of klassen. Het meten is dan het toewijzen van een persoon aan een bepaalde klasse, terwijl de klassen onderling niet met elkaar in verband worden gebracht. Het model op zichzelf is uiterst eenvoudig. Stel dat er $A$ klassen zijn. De conditionele kans op een antwoordpatroon $\boldsymbol{x}$, gegeven dat het afkomstig is van een persoon uit klasse $a$ is gegeven door

$$\pi_{\boldsymbol{x} \mid a} = \pi_{x_1 \mid a} \pi_{x_2 \mid a} \ldots \pi_{x_k \mid a}, \quad (5.17)$$

waarin men direct een toepassing herkent van het principe van de lokale stochastische onafhankelijkheid. De data bestaan echter uit de antwoordpatronen $\boldsymbol{x}$ en het klasselidmaatschap van een persoon is niet geobserveerd. Als de kans dat een persoon

behoort tot klasse $a$ voorgesteld wordt door $\pi_a$, $(a = 1,..., A)$, is de marginale kans op een antwoordpatroon $\boldsymbol{x}$ gegeven door

$$P(\boldsymbol{x}) = \sum_a \pi_{\boldsymbol{x}|a}\pi_a = \sum_a \pi_{x_1|a}...\pi_{x_k|a}\pi_a. \tag{5.18}$$

In het geval van dichotome items moet dus voor elk item de conditionele kans op een antwoord geschat worden gegeven de klasse $a$, $\pi_{x_i|a}$, en daarenboven moeten $A$-1 onafhankelijke kansen $\pi_a$ geschat worden. Hoewel het model op zichzelf een heel eenvoudige structuur heeft, is de schatting van de parameters geen triviaal probleem. Dit model is een van de eerste IRT-modellen en werd voorgesteld door Lazarsfeld (1950). Het model kreeg van Lazarsfeld de naam latente-klassenmodel, omdat het klasselidmaatschap niet geob-serveerd, dus latent is. Lazarsfeld gebruikte trouwens niet het begrip IRT maar de algemene benaming 'Latente-structuuranalyse' om modellen met latente variabelen aan te duiden.

<table>
<tr><td colspan="3">Monotone items</td><td>Niet-monotone items</td></tr>
<tr><td rowspan="2">Unidimensionaal</td><td>Dichotoom</td><td>Hoofdst. 4 en 5.3</td><td rowspan="3">Ontvouwings modellen</td></tr>
<tr><td>Polytoom</td><td>5.4</td></tr>
<tr><td>Multidimensionaal</td><td>Dichotoom en polytoom</td><td>5.5</td></tr>
<tr><td>A-dimensionaal</td><td colspan="3">Latente-klassenmodellen</td></tr>
</table>

Figuur 5.3
Een indeling van itemresponsmodellen

In figuur 5.3 is een schematische weergave gegeven van de indeling van IRT-modellen die hiervoor werd besproken. De gearceerde oppervlakken bevatten een verwijzing naar de paragrafen in dit hoofdstuk waar een of meer modellen uit de cel van de figuur zullen worden besproken.

Het valt in figuur 5.3 op dat het onderscheid in monotone en niet-monotone items niet gehandhaafd is bij a-dimensionale gevallen. Dit kan ook niet anders, want het begrip monotoniciteit heeft geen enkele betekenis als de waarden van de latente variabele niet geordend kunnen worden. De indeling van IRT-modellen als in figuur

5.3 is voorgesteld is zeker niet de enig mogelijke. Ze is bedoeld als een handvat om enige orde te scheppen in de grote hoeveelheid modellen die in de literatuur zijn beschreven. Andere indelingen, die ook andere verbanden duidelijker belichten, zijn gegeven door Masters en Wright (1984), Thissen en Steinberg (1986) en Heinen (1993).

## 5.3 Unidimensionale modellen voor dichotome items

In hoofdstuk 4 is op verschillende plaatsen gewezen op een paar kwetsbare punten van het Raschmodel, namelijk de strenge eis dat alle items gelijkelijk moeten discrimineren en het feit dat het Raschmodel ongeschikt is om de relatief grote kansen op een juist antwoord te verklaren wanneer er geraden wordt bij meerkeuze-items. In de literatuur zijn modellen ontwikkeld die op het eerste gezicht een afdoend antwoord bieden op deze problemen. De meest prominente modellen zijn het twee- en het drieparameter logistisch model. Deze twee modellen worden besproken in paragraaf 5.3.1. We zullen echter zien dat het gebruik van deze modellen niet helemaal zonder problemen is omdat hierbij bepaalde aantrekkelijke eigenschappen van het Raschmodel verloren. Met name de mogelijkheid om itemparameters met de CML-methode te schatten is niet meer aanwezig. In paragraaf 5.3.2 wordt een model besproken dat de flexibiliteit van het tweeparameter logistisch model koppelt aan de theoretische voordelen van het Raschmodel. Het is het zogenaamde éénparameter logistisch model (Engels: One Parameter Logistic Model, OPLM).

In paragraaf 5.3.3 wordt ingegaan op modellen die geschikt zijn wanneer het axioma van de lokale stochastische onafhankelijkheid geschonden is. Te zelfder tijd zullen we zien dat het gebruik van deze modellen, in samenhang met de constructie van LR-toetsen, toelaat de geldigheid van dit axioma statistisch te toetsen.

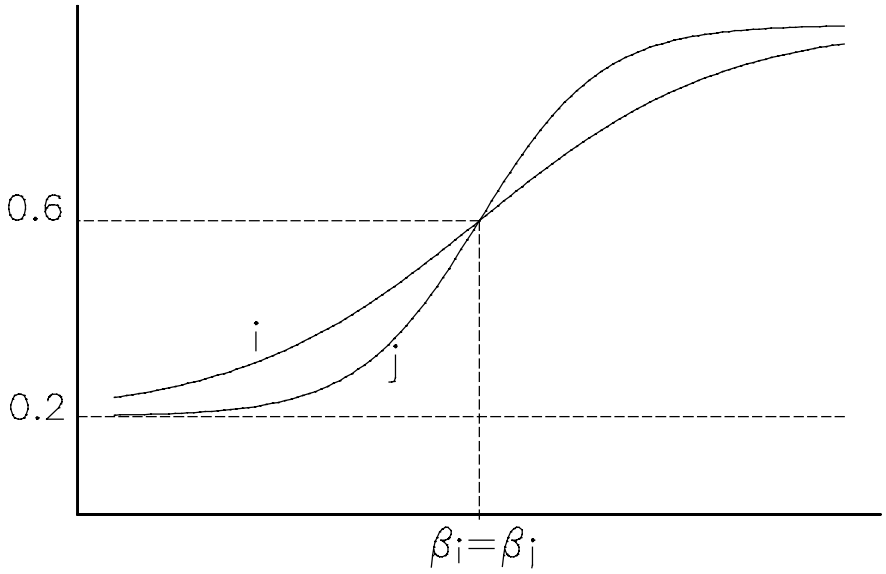
### *5.3.1 Het twee- en drieparameter logistisch model*

Het tweeparameter logistisch model (Birnbaum, 1968) werd reeds kort besproken in hoofdstuk 4. Hier beginnen we met het drieparameter logistisch model dat eveneens door Birnbaum (1968) is beschreven. Een uitvoerige discussie over dit model kan men vinden in Lord (1980). Daarna zullen we zien dat het tweeparametermodel beschouwd kan worden als een speciaal geval van het drieparametermodel. In de literatuur worden

deze modellen vaak afgekort met 2PL en 3PL, deze afkortingen zullen we ook hier gebruiken. De itemresponsfunctie in het 3PL is gegeven door:

$$f_i(\theta) = c_i + (1 - c_i)\frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}, \quad (a_i > 0\ ;\ 0 \leq c_i < 1). \tag{5.19}$$

In figuur 5.4 staan de grafieken van twee itemresponsfuncties $f_i(\theta)$ en $f_j(\theta)$ met $\beta_i = \beta_j$, $c_i = c_j = 0.2$, $a_i = 1$ en $a_j = 2$.




Figuur 5.4
Itemresponsfuncties in het 3PL

De curve van item *j* verloopt steiler dan die van item *i*, hetgeen het effect van een grotere discriminatieparameter weerspiegeld. Het is gemakkelijk na te gaan dat in het 3PL de volgende limieten gelden

$$\lim_{\theta \to \infty} f_i(\theta) = 1$$

$$\lim_{\theta \to -\infty} f_i(\theta) = c_i$$

De parameter $c_i$ geeft dus de kans op een juist antwoord aan indien de vaardigheid zeer klein is. Iets losser geformuleerd zou men kunnen zeggen dat $c_i$ de kans is op een juist antwoord als men het antwoord niet 'kent'. Dit model lijkt dus geknipt te zijn voor toepassing bij meerkeuze-vragen. De parameter $c_i$ wordt dan ook vaak aangeduid als de raadparameter. De interpretatie van deze parameter is echter ingewikkelder dan het op het eerste gezicht lijkt. In de eerste plaats is het 3PL uitsluitend gedefinieerd door

(5.19) en de bijkomende aanname van lokale stochastische onafhankelijkheid. De interpretatie van $c_i$ als raadparameter maakt geen deel uit van het model. Indien we data hebben die uitstekend beschreven worden door het 3PL, volgt daar niet logisch uit dat er geraden is. Het zou bijvoorbeeld zo kunnen zijn dat personen die het juiste antwoord niet echt kennen, toch een of andere, verkeerde, redenering volgen die met een kans $c_i$ in het juiste antwoord resulteert. Het is nuttig om na te gaan of we niet een model van het cognitieve functioneren kunnen opstellen dat dezelfde voorspellingen maakt als het 3PL. Daartoe definiëren we een nieuwe functie die we zullen aanduiden met het symbool $h_i$:

$$h_i(\theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}. \qquad (5.20)$$

De functie $h_i(\theta)$ komt dus overeen met de breuk in het rechterlid van (5.19). Het is duidelijk dat $0 < h_i(\theta) < 1$. We interpreteren deze functie als de kans dat een persoon met vaardigheid $\theta$ het antwoord op het item kent. Voorts veronderstellen we dat, indien het juiste antwoord 'geweten' wordt, het ook daadwerkelijk gegeven wordt. Dat wil zeggen dat we hier aannemen dat de persoon zich niet kan vergissen, later zullen we onderzoeken wat er gebeurt als we deze assumptie laten vallen. Indien de persoon het antwoord niet kent, wordt er met een kans $1 - h_i(\theta)$ overgegaan op raden en het juiste antwoord wordt dan geraden met kans $c_i$. De verschillende gebeurtenissen en kansen zijn schematisch weergegeven in tabel 5.1.

Tabel 5.1
Een cognitief model voor het beantwoorden van meerkeuze-items

| Gebeurtenis | Kans | Antwoord |
|---|---|---|
| Kent antwoord en vergist zich niet | $1 \times h_i(\theta) = h_i(\theta)$ | Juist |
| Kent antwoord maar vergist zich | $0 \times h_i(\theta) = 0$ | Fout |
| Kent antwoord niet maar raadt juist | $c_i \times [1 - h_i(\theta)]$ | Juist |
| Kent antwoord niet en raadt verkeerd | $(1 - c_i) \times [1 - h_i(\theta)]$ | Fout |

De kans op een juist antwoord is dus de som van de twee kansen uit tabel 5.1 die tot een juist antwoord leiden:

$$P(X_i = 1 \mid \theta) = h_i(\theta) + c_i[1 - h_i(\theta)]$$
$$= c_i + (1 - c_i)\, h_i(\theta) = f_i(\theta).$$

Het eenvoudige cognitieve model leidt dus tot het 3PL. Binnen dit cognitieve model kunnen we dan ook de kans berekenen dat een juist antwoord door raden tot stand is gekomen:

$$P(\text{raden} \mid X_i = 1, \theta) = \frac{c_i[1 - h_i(\theta)]}{h_i(\theta) + c_i[1 - h_i(\theta)]}. \tag{5.21}$$

Het rechterlid van (5.21) is niet te vereenvoudigen, omdat de afhankelijkheid van θ er in aanwezig blijft. Dit betekent dat we geen nauwkeurige uitspraak kunnen doen over de hoeveelheid juiste antwoorden die door raden tot stand zijn gekomen in een willekeurige steekproef van antwoordpatronen. We kunnen het wel indien we de verdeling van θ kennen. Indien $g(\theta)$ de dichtheidsfunctie is van θ vinden we:

$$P(\text{raden} \mid X_i = 1) = \int_{-\infty}^{\infty} \frac{c_i[1 - h_i(\theta)]}{h_i(\theta) + c_i[1 - h_i(\theta)]}\, g(\theta)\, d\theta. \tag{5.22}$$

De dichtheidsfunctie $g(\theta)$ maakt echter geen deel uit van het 3PL, maar moet er aan toegevoegd worden.

Samenvattend kunnen we zeggen dat het cognitieve model, in de mate dat het een min of meer realistische voorstelling van cognitieve processen geeft, een rechtvaardiging is van het 3PL, maar dat het niet door het 3PL wordt geïmpliceerd. We keren nu terug naar een verdere analyse van het 3PL.

In het Raschmodel hebben we de moeilijkheidsgraad van een item omschreven als de hoeveelheid vaardigheid die nodig is om een kans te hebben van precies 0.5 om het item juist te beantwoorden. Deze interpretatie van de itemparameter geldt niet meer in het 3PL. Indien θ gelijk is aan $\beta_i$ krijgen we

$$f_i(\beta_i) = c_i + (1 - c_i) \times 0.5 = 0.5 + \frac{c_i}{2}. \tag{5.23}$$

De interpretatie van $\beta_i$ als moeilijkheidsparameter is dus niet zo overtuigend als in het Raschmodel, door de afhankelijkheid van $c_i$ die in (5.23) tot uiting komt. Toch wordt de parameter $\beta_i$ in de literatuur aangeduid als moeilijkheidsparameter.

Wellicht ten overvloede vermelden we nog even dat het model (5.19) niet geïdentificeerd is. Het linkerlid van (5.19) verandert niet als bij de β-parameters en bij θ een willekeurige constante $c$ wordt opgeteld. Het nulpunt van de schaal kan dus, net

als bij het Raschmodel, vrij gekozen worden. Bovendien kunnen we zowel $\theta$ als $\beta_i$ met een willekeurige positieve constante vermenigvuldigen, als we te zelfder tijd $a_i$ door die constante delen. Dit betekent dat we de eenheid van de schaal willekeurig kunnen kiezen. Die keuze kunnen we bijvoorbeeld maken door te eisen dat $a_1 = 1$. De parameters $c_i$ liggen op een absolute schaal en kunnen niet getransformeerd worden.

Tenslotte nog een terminologische kwestie. Het rechterlid van (5.19) kan niet teruggebracht worden tot de standaardvorm van de logistische functie. Strikt genomen is het 3PL dus geen logistisch model, maar in de literatuur wordt het wel zo genoemd. Wij zullen ons aan dit gebruik conformeren.

Het 2PL kan men opvatten als een speciaal geval van het 3PL: het is gegeven door in (5.19) de parameter $c_i$ gelijk te stellen aan 0 voor alle items. De itemresponsfunctie in het 2PL valt dus samen met de functie $h_i(\theta)$ die in (5.20) is gedefinieerd. Wanneer we verderop het 2PL onderzoeken, zullen we echter niet het functiesymbool $h$ gebruiken maar $f$ om de itemresponsfunctie aan te duiden.

### *Parameterschatting in het 2PL en het 3PL*

Bij een eerste beschouwing van (5.19) zou men de volgende redenering kunnen volgen. Het 2PL is een speciaal geval van het 3PL en het Raschmodel is op zijn beurt weer een speciaal geval van het 2PL, dat ontstaat door alle discriminatieparameters aan elkaar gelijk te stellen. Als we dus altijd werken met het 3PL, merken we vanzelf wel of de raadparameters gelijk zijn aan 0 of niet en of de discriminatieparameters gelijk zijn of ongelijk. De realiteit is niet zo eenvoudig. Het schatten van de parameters in het 2PL en het 3PL is namelijk heel wat moeilijker dan in het Raschmodel en bovendien is het uitmaken of het 2PL of het Raschmodel passende modellen zijn niet eenvoudig. Om deze moeilijkheden te illustreren zullen we ons in eerste instantie beperken tot het 2PL. Later zullen we nog enkele beschouwingen toevoegen over het 3PL.

De log-aannemelijkheidsfunctie gegeven een antwoordpatroon $\boldsymbol{x}$ voor het 2PL werd reeds besproken in hoofdstuk 4, formule (4.61). We herhalen deze formule hier:

$$\ln L(\beta, \boldsymbol{a}, \theta; \boldsymbol{x}) = \theta \sum_i a_i x_i - \sum_i x_i a_i \beta_i - \sum_i \ln\{1 + \exp[a_i(\theta - \beta_i]\}. \qquad (5.24)$$

Het is direct duidelijk dat CML als schattingsprocedure is uitgesloten. We kunnen niet conditioneren op $\sum_i a_i x_i$ omdat deze grootheid afhankelijk is van de onbekende parameters $a_i$. Van de schattingsmethoden die in hoofdstuk 4 werden besproken, blijven dus alleen JML en MML over. Bij de JML-methode hebben we een analoog probleem

als bij het Raschmodel. Door de aanwezigheid van de incidentele parameters $\theta_v$ kunnen we geen beroep doen op standaardresultaten uit de statistiek. Met name weten we niet of de JML-schatters wel consistent zijn. Het is niet zo dat de aanwezigheid van incidentele parameters in alle gevallen leidt tot inconsistentie van de schatters van de structurele parameters, maar als er incidentele parameters zijn en men wil toch gebruik maken van JML, dan dient men de consistentie van de schatters aan te tonen. Een dergelijk bewijs voor het 2PL is in de IRT-literatuur echter nog nooit gegeven. Hierna geven wij de schets van een bewijs dat JML in het 2PL geen consistente schatters oplevert van de β-parameters en ook niet van de discriminatieparameters. We doen dit aan de hand van het eenvoudigst mogelijke geval met $k = 2$ items.

Bij twee items zijn er maar vier mogelijke antwoordpatronen: (0 0), (0 1), (1 0) en (1 1). Bij een steekproef van $n$ personen kunnen we de observaties dus handig samenvatten door de frequenties van die vier antwoordpatronen te hanteren. Deze frequenties worden aangeduid als respectievelijk $n_{00}, n_{01}, n_{10}$ en $n_{11}$. Het aantal itemparameters dat in het 2PL moet worden geschat is $2(k-1)$, $k-1$ β-parameters en $k-1$ discriminatieparameters. Omdat we met JML werken en dus met elke persoon een parameter associëren, moeten bovendien nog $n$ persoonsparameters geschat worden. We kiezen de normering van de schaal zo dat $\beta_1 = 0$ en $a_1 = 1$. We moeten dus $\beta_2, a_2, \theta_1, ..., \theta_n$ schatten. De schattingen kunnen we met standaardtechnieken berekenen, door de partiële afgeleiden van de log-aannemelijkheidsfunctie gelijk te stellen aan 0 en de aldus ontstane vergelijkingen op te lossen. Voor het geval $k = 2$ kan een expliciete oplossing gevonden worden. We zullen de details van de afleiding niet bespreken, maar geven alleen het resultaat. Daarbij veronderstellen we dat $n_{01}$ en $n_{10}$ beide van 0 verschillen.

(1) Personen met hetzelfde antwoordpatroon krijgen dezelfde schatting van θ. De schattingen van de $n$ θ-parameters kunnen dus niet meer dan vier verschillende waarden aannemen, die we zullen aan duiden als $\hat{\theta}_{00}$, $\hat{\theta}_{01}$, $\hat{\theta}_{10}$ en $\hat{\theta}_{11}$.

(2) $\hat{\theta}_{00}$ en $\hat{\theta}_{11}$ bestaan niet. Dit wil zeggen dat er geen reële getallen bestaan die we voor die twee schatters kunnen invullen zodat aan de schattingsvergelijkingen is voldaan. Dit impliceert eigenlijk dat we het probleem iets anders moeten formuleren en zeggen dat we onze schattingen gaan baseren op de $n_{01} + n_{10}$ antwoordpatronen die precies één item juist hebben.

(3) $\hat{\theta}_{01} = \hat{\theta}_{10} = \ln(n_{10}/n_{01})$, dus alle personen met één juist antwoord krijgen dezelfde schatting van θ.

(4) $\hat{a}_2 = 1$, of iets algemener gezegd, $a_2$ wordt geschat op precies dezelfde waarde die we aan $a_1$ hebben toegekend.

(5) $\hat{\beta}_2 = 2\ln(n_{10}/n_{01})$.

Uit resultaat (4) volgt direct dat de discriminatieparameters niet consistent geschat worden: wat ook de steekproefomvang is en wat de echte waarden van de discriminatieparameters ook zijn, ze worden steeds als even groot geschat. Om de inconsistentie van de schatter van $\beta_2$ aan te tonen, beschouwen we een speciaal geval van het 2PL waar de discriminatieparameters aan elkaar gelijk zijn. Dan krijgen we voor $\beta_2$ natuurlijk dezelfde schatter die in resultaat (5) is gegeven. Maar dit speciale geval van het 2PL is niets anders dan het Raschmodel en de schatter in (5) is ook precies dezelfde als de JML-schatter van $\beta_2$ in het Raschmodel (Fischer, 1974, p. 260), waarvan is aangetoond dat hij inconsistent is. Het besluit is dus dat de itemparameters in het 2PL niet consistent geschat worden. Dit resultaat sluit niet uit dat de schatters bij een andere *k* misschien wel consistent zijn, doch dit zou dan moeten worden aangetoond.

Het niet consistent zijn van schatters heeft grote gevolgen voor de toepassingen van een model. Losweg betekent het niet-consistent zijn, dat de schattingen systematisch gaan afwijken van de werkelijke waarden en dat die systematische fout niet verholpen kan worden door de steekproef groter te maken. Dit hoeft in bepaalde opzichten niet erg te zijn. Als de systematische fout klein is, zouden we daar genoegen mee kunnen nemen. Zo blijkt in het Raschmodel bijvoorbeeld, dat de systematische fout kleiner wordt als *k* toeneemt. Bovendien kan men in het Raschmodel een correctie aanbrengen op de JML-schattingen door ze te vermenigvuldigen met $(k-1)/k$. Uit simulatiestudies blijkt dat de aldus gecorrigeerde JML-schattingen erg goed overeenkomen met de CML-schattingen die wel consistent zijn. Dit is een nuttig resultaat, maar het lost slechts een deelprobleem op. Alle theorie die in hoofdstuk 4 is behandeld over standaardfouten en de asymptotische verdeling van toetsingsgrootheden, is niet zonder meer geldig in het geval dat de ML-schatters niet consistent zijn. Men kan natuurlijk in een concrete toepassing de geobserveerde informatiematrix inverteren en de elementen op de diagonaal beschouwen als schatters van de variantie, doch men kent niet meer de eigenschappen van die schatters en die zouden wel eens erg onaantrekkelijk kunnen zijn. Het feit dat er veel publikaties zijn in de IRT-literatuur waar deze procedure wordt toegepast, kan niets veranderen aan het dubieuze karakter ervan.

Het gebruik van de MML-procedure omzeilt de problemen van de incidentele parameters. Zoals in hoofdstuk 4 reeds is benadrukt, dient men echter wel te bedenken dat MML niet alleen een procedure is, maar dat het meetmodel uitgebreid wordt met een veronderstelling over de verdeling van $\theta$. Verder is de uiteenzetting over MML uit

hoofdstuk 4 ook van toepassing op het 2PL en het 3PL. Op de problemen van algoritmische en numerieke aard gaan we hier niet verder in. Gedetailleerde uiteenzettingen hierover kan men vinden in Bock en Aitkin (1981) en in Rigdon en Tsutakawa (1983).

Er is echter één probleem dat ogenschijnlijk veel te maken heeft met de berekening van de schattingen, maar dat een veel diepere oorzaak heeft die te maken heeft met de eigenschappen van het model. We kunnen het probleem het beste illustreren aan de hand van het 3PL. Indien we het Raschmodel toepassen, vinden we altijd dat een item met een grote *p*-waarde een kleinere geschatte moeilijkheidsparameter heeft dan een item met een kleine *p*-waarde. Men kan aantonen dat dit mathematisch noodzakelijk is, en het is ook wat we normaliter zouden verwachten. Bij het 3PL verschijnt echter een dubbelzinnigheid: een grote *p*-waarde kan wijzen op een gemakkelijk item en een kleine raadparameter maar ook op een moeilijk item met een grote raadparameter. De itemantwoorden zijn dus in zekere zin dubbelzinnig: uit de kwaliteit van het antwoord kan men de waarde van de parameters moeilijk afleiden. Of anders gezegd, de data bevatten erg weinig informatie die gebruikt kan worden om onderscheid te maken tussen moeilijkheid en raadkans. Dit heeft tot gevolg dat het vinden van het maximum van de aannemelijkheidsfunctie in het algemeen moeilijker zal zijn dan in het Raschmodel en dat de nauwkeurigheid waarmee de parameters geschat worden kleiner zal. Bovendien ontspoort de schattingsprocedure soms door een oplossing op te leveren die niet overeenkomt met het maximum van de aannemelijkheidsfunctie. Als item *i* een vierkeuze-item is, ver-wachten we dat de schatting van $c_i$ niet al te ver zal afwijken van 0.25. Krijgen we als resultaat echter een schatting van 0.85, dan zullen we niet al te snel geneigd zijn met deze schatting genoegen te nemen. Deze problemen ontstaan dus eigenlijk omdat we de data overvragen, of vanuit een ander standpunt bekeken, omdat we te weinig informatie hebben verzameld. Indien we een betrouwbare procedure konden verzinnen waarbij de persoon bij elk itemantwoord ook aangeeft of er geraden is of niet, dan zouden we veel meer informatie hebben en we zouden ook veel nauwkeuriger kunnen schatten.

De voorgaande beschouwing geeft ook aan dat er in zekere zin grenzen zijn aan de complexiteit van IRT-modellen. Het is niet moeilijk om het cognitieve model dat in tabel 5.1 is weergegeven iets realistischer te maken, door de kans op een vergissing als men het antwoord kent niet gelijk te stellen aan 1, maar daar een nieuwe parameter $d_i$ voor te kiezen. Dit leidt dan tot een 4PL, waarvan het in principe mogelijk is de parameters te schatten als men alleen over dichotome data beschikt. De schattingen zullen echter zo instabiel zijn dat ze in de praktijk eigenlijk niet meer bruikbaar zijn, tenzij men over gigantische steekproeven kan beschikken.

Er bestaat echter ook een andere manier om het tekort aan informatie te ondervangen, namelijk het toepassen van een schattingstechniek die afkomstig is uit de bayesiaanse statistiek. Hier voegt men zijn ongeloof dat de *c*-parameter uit het voorbeeld gelijk is aan 0.85 op een formele manier aan het model toe door middel van een a priori verdeling, die voor alle mogelijke waarden van de parameter als het ware de voorafgaande overtuiging uitdrukt dat de parameter die waarde aanneemt. Als de a priori verdeling uniform is, drukken we daarmee uit dat we eigenlijk helemaal niets weten over die parameter. Is die verdeling eentoppig met een hele kleine standaardafwijking en met modus of gemiddelde in de buurt van 0.25, dan geven we daarmee aan dat we er vrijwel zeker van zijn dat de raadkans niet ver van 0.25 zal afwijken. De observaties worden dan gebruikt om onze overtuiging te wijzigen: de gegevens en de a priori verdeling worden met elkaar gecombineerd en leveren een nieuwe verdeling van de parameter op die de a posteriori verdeling genoemd wordt en die op haar beurt weer kan fungeren als a priori verdeling voor toekomstige observaties. Als schatter van de parameter neemt men dan een of ander kenmerk van de a posteriori verdeling, zoals de modus of het gemiddelde en als maat van onzekerheid neemt men meestal de standaardafwijking van de a posteriori verdeling. Een meer technische uiteenzetting is gegeven in paragraaf 4.5 bij de behandeling van de EAP-schatter van $\theta$ in het Raschmodel. Men kan deze techniek ook toepassen bij meer parameters tegelijk, maar dan moet men een a priori verdeling specificeren voor alle parameters tegelijk. In dat geval blijkt het berekenen van de modus van de multivariate a posteriori verdeling meestal eenvoudiger te zijn dan het berekenen van het gemiddelde. Deze techniek wordt bijvoorbeeld toegepast in het computerprogramma BILOG (Mislevy & Bock, 1986) dat de parameters voor het 3PL, het 2PL en het Raschmodel schat en dat in de regel plausibele schattingen oplevert.

Hoewel het gebruiken van een bayesiaanse benadering erg elegant is en veel problemen van JML en MML omzeilt, dient men toch de nodige voorzichtigheid in acht te nemen bij het gebruik van deze techniek. Op het eerste gezicht lijkt deze benadering een element van willekeur te bevatten. Iedereen kan immers zijn eigen a priori verdeling kiezen, waardoor ook steeds, bij dezelfde data, verschillende schattingen zullen worden verkregen. De wetenschappelijke consensus zal zo ver te zoeken zijn. De bayesiaanse statistiek heeft een adequaat antwoord op dit bezwaar. Ten eerste moet de rol van de a priori verdeling niet overschat worden. Indien er maar voldoende observaties zijn, wordt de a posteriori verdeling bijna volledig bepaald door de observaties en speelt de a priori verdeling geen rol van betekenis meer. Ten tweede is de a priori verdeling bedoeld als een soort samenvatting van eerder gedane observaties en ervaringen. Als twee onderzoekers in hetzelfde domein van wetenschap actief zijn,

dezelfde literatuur lezen en vergelijkbaar onderzoek doen, kunnen hun overtuigingen in de bayesiaanse betekenis niet drastisch van elkaar verschillen. Maar dat is theorie. In de praktijk kan de misvatting optreden dat het er niet toe doet welke a priori verdeling men kiest, omdat het aantal van 200 observaties waarover men beschikt geweldig groot is vergeleken met de 25 waarop de collega of de concurrent zijn analyse uitvoerde. Of een steekproef groot genoeg is om de a priori verdeling onbelangrijk te maken, hangt af van de standaardafwijking van de a priori verdeling. Kiest men deze standaardafwijking erg klein, dan kan bij een steekproef die gevoelsmatig erg groot lijkt, de a posteriori modus zeer dicht bij de modus van de a priori verdeling liggen. Als bewijs dat men het met de a priori verdeling 'dus' bij het rechte eind had, is dit echter niet overtuigend. Men heeft bij wijze van spreken aangetoond dat men zo'n sterke overtuiging had, dat die door de 100 of 200 observaties waarover men beschikt niet wezenlijk te veranderen is. Kiest men de standaardafwijking echter te groot, dan is de a posteriori verdeling grotendeels bepaald door de observaties en gaat de schattingsprocedure erg lijken op de ML-schattingsprocedure en verliest de bayesiaanse benadering eigenlijk haar zin.

### *Statistische toetsen voor het 2PL en het 3PL*

De behandeling van dit onderwerp kan kort zijn, om de eenvoudige reden dat er zeer weinig toetsen zijn ontwikkeld die voor deze modellen gebruikt kunnen worden. Waarom dit zo is, is niet gemakkelijk te zeggen, doch we kunnen zeker twee mogelijke redenen aangeven. De eerste reden heeft te maken met de moeilijkheid van het probleem. Alles wat in hoofdstuk 4 is gezegd over het construeren van veralgemeende $X^2$-toetsen had betrekking op modellen uit de exponentiële familie. Het 2PL en het 3PL behoren niet tot deze familie. Glas (1989) heeft weliswaar aangetoond dat er gelijkaardige toetsen geconstrueerd kunnen worden voor modellen buiten de exponentiële familie, zoals de $R_0$- en de $R_{1m}$-toetsen, maar de bewijsvoering is heel specifiek voor het Raschmodel en is niet zonder meer bruikbaar voor het 2PL en het 3PL.

De tweede reden heeft te maken met een verschil van instelling tussen de Europese psychometrici enerzijds en een groot gedeelte van de Amerikaanse vakgenoten. De Europese literatuur over IRT is zeer sterk beïnvloed door het werk van Rasch (1960) en Fischer (1974), waar een grote nadruk gelegd wordt op de theoretische eigenschappen die in een deugdelijk meetinstrument aanwezig moeten zijn. Dit heeft niet alleen geleid tot de prominente plaats die het Raschmodel in de IRT-literatuur

inneemt, maar ook tot een grote inspanning om statistische toetsen te ontwerpen waarmee kan worden nagegaan of aan de strenge eisen van het Raschmodel is voldaan. De Amerikaanse literatuur over IRT daarentegen is zeer sterk beïnvloed door het werk van F. Lord, die gezien zijn werkzaamheden op het toetsinstituut Educational Testing Service (ETS) een veel pragmatischer instelling had. Waar men het devies van de Europese traditie grofweg zou kunnen omschrijven als: 'maak toetsen die aan het Raschmodel voldoen', kwam Lords devies neer op: 'maak modellen die adequaat zijn voor de bestaande toetsen'. Door het wijdverspreide gebruik van meerkeuze-items is de ontwikkeling en het gebruik van het 3PL dan ook goed te begrijpen. Omdat dit model voorziet in verschillende discriminatieparameters voor de items en in een onderste asymptoot die verschillend kan zijn van 0, is er ook minder behoefte aan statistische toetsing. De twee voor de hand liggende kwetsbare plekken van het Raschmodel zijn immers modelmatig weggewerkt.

Het hierboven geschetste verschil in benadering van de IRT is natuurlijk niet absoluut en er zijn statistische toetsen ontwikkeld die van toepassing zijn voor het 2PL en het 3PL. Deze toetsen zijn besproken in paragraaf 4.3.5 als varianten van de $S_i$-toetsen. Bovendien is het natuurlijk mogelijk LR-toetsen te construeren waarin het 2PL of het 3PL als nulhypothese fungeert en het verzadigde multinomiale model als alternatieve hypothese. Men zou kunnen opperen dat een LR-toets waarbij het 2PL fungeert als nulhypothese en het 3PL als algemeen model of alternatieve hypothese meer onderscheidingsvermogen zal hebben. Dit is echter geen goed idee. Bij de bespreking van de LR-toetsen in hoofdstuk 4 hebben we gezien dat bij een LR-toets de parameterruimte van het beperkte model een deelruimte moet zijn van de parameterruimte in het algemene model. De eis is echter strenger. De beperkte parameter-ruimte moet helemaal binnen de algemene parameterruimte liggen. We gaan hier niet in op de precieze mathematische betekenis van 'binnen', maar we illustreren het principe met een voorbeeld. Als we het 2PL beschouwen als een speciaal geval van het 3PL, betekent dit dat we alle $c_i$-parameters in het 3PL fixeren op de waarde 0, maar deze waarde is de kleinste waarde die de $c_i$-parameters kunnen aannemen. Men zegt dat de parameters in het 2PL gefixeerd worden op de rand van de parameterruimte van het 3PL en in dit geval mag men zeker niet zonder meer aannemen dat de LR-toetsingsgrootheid chi-kwadraat verdeeld is.

### *5.3.2 Het éénparameter logistisch model (OPLM)*

Er zijn vele varianten mogelijk op het 3PL, waarvan sommige als gevolg van moeilijkheden bij het schatten van de parameters in het algemene 3PL daadwerkelijk in de literatuur zijn toegepast. Meestal gaat het om beperkingen op de $c_i$-parameters. Indien in een meerkeuzetoets alle items evenveel antwoordalternatieven hebben, zou men het redelijk kunnen vinden te eisen dat alle $c_i$-parameters aan elkaar gelijk zijn. Deze eis komt overeen met het opleggen van $k$-1 lineaire restricties aan de parameters van het model, analoog aan wat gebeurt bij de moeilijkheidsparameters in het LLTM. Een verdere restrictie die soms wordt toegepast, bestaat erin die gemeenschappelijke $c$-parameter gelijk te stellen aan één gedeeld door het aantal antwoordalternatieven. Door deze eis verandert de status van $c$. Het is geen onbekende grootheid meer die uit de data moet worden geschat, maar een bekende constante. Hoewel deze twee varianten van het 3PL het schattingsprobleem sterk vereenvoudigen, is er geen mogelijkheid om CML toe te passen.

Er bestaat echter wel een mogelijkheid om dusdanige restricties op het 2PL aan te brengen dat CML wel mogelijk wordt. Indien we in (5.24) de grootheden $a_i$ niet langer beschouwen als onbekende parameters maar als gegeven constanten, zien we dat deze speciale versie van het 2PL tot de exponentiële familie behoort en dat de gewogen score $s = \sum_i a_i x_i$ een groot- heid is die zonder meer uit de data kan worden berekend en waarop dus geconditioneerd kan worden. Hierdoor verliest $a_i$ zijn status van parameter. Om dit essentiële onderscheid in de terminologie goed aan te geven, zullen we spreken van discriminatie-indices. Het model werd voorgesteld door Verhelst en Eggen (1989) en kreeg de naam éénparameter logistisch model (OPLM) op grond van het argument dat er per item slechts één parameter overblijft.

Bij de bespreking van het 2PL hebben we gezien dat één discriminatieparameter vrij gekozen kan worden en dat daarmee de eenheid van de schaal wordt vastgelegd. Welke waarde we kiezen doet niet ter zake. Bijgevolg is een uitspraak als: ’dit item discrimineert erg goed want zijn discriminatieparameter is gelijk aan 5’ zinloos als niet, expliciet of impliciet, gerefereerd wordt naar de eenheid van de schaal. Deze referentie is altijd aanwezig indien men verhoudingen van discriminatieparameters of -indices hanteert. Dit maakt ook duidelijk dat, indien alle discriminatie-indices met een constante worden vermenigvuldigd, het model niet verandert. Nu kunnen we die constante zo kiezen dat de resulterende indices allemaal gehele getallen zijn of willekeurig dicht door een geheel getal kunnen worden benaderd. Het houdt dus nauwelijks een beperking in als we zeggen dat de discriminatie-indices gehele getallen moeten zijn. In de verdere bespreking zullen we daar dan ook van uitgaan. Merk op dat het Raschmodel een speciaal geval is van het OPLM, waarin alle discriminatie-indices aan elkaar gelijk zijn.

Met betrekking tot de schatting van de itemparameters in het OPLM hoeven we nauwelijks iets toe te voegen aan de discussie die in hoofdstuk 4 is gewijd aan de parameterschattingen in het Raschmodel. Door een geschikte parametrisering te kiezen, blijken de formules die we gebruikt hebben bij de bespreking van het Raschmodel formeel gelijk te zijn aan de formules voor het OPLM. De conditionele aannemelijkheidsfunctie kan dus geschreven worden als:

$$\ln L(\boldsymbol{\varepsilon}; X \mid \boldsymbol{s}) = \sum_i t_i \ln \varepsilon_i - \sum_v \ln \gamma_{s_v}(\boldsymbol{\varepsilon}), \tag{5.25}$$

en die formule is precies gelijk aan (4.43). Alleen is de parameter $\varepsilon_i$ nu gedefinieerd als

$$\varepsilon_i = \exp(-a_i \beta_i). \tag{5.26}$$

Merk op dat met $s_v$ de gewogen score bedoeld wordt en met $t_i = \sum_v x_{vi}$ het aantal juiste antwoorden dat op item *i* is uitgebracht. De functie $\gamma_s(\boldsymbol{\varepsilon})$ is formeel gedefinieerd als

$$\gamma_s(\boldsymbol{\varepsilon}) = \sum_{\sum a_i x_i = s} \prod_i \varepsilon_i^{x_i}. \tag{5.27}$$

We geven een voorbeeld om de structuur van (5.27) te verduidelijken. Veronderstel dat $k = 4$ en de eerste drie items een discriminatie-index gelijk aan 1 hebben, maar dat $a_4 = 2$. Er zijn precies vier antwoordpatronen die een gewogen score van 2 opleveren: (1 1 0 0), (1 0 1 0), (0 1 1 0) en (0 0 0 1). De som die we nodig hebben om $\gamma_2(\boldsymbol{\varepsilon})$ uit te rekenen zal bijgevolg uit vier termen bestaan:

$$\gamma_2(\boldsymbol{\varepsilon}) = \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3 + \varepsilon_4.$$

In tegenstelling tot de symmetrische functies die we nodig hadden bij het Raschmodel, komen in het rechterlid van bovenstaande uitdrukking niet meer alle tweetallen van parameters voor als produkt, maar alleen die combinaties van parameters die overeenkomen met een gewogen score van 2. De $\gamma$-functies zijn dus niet langer symmetrisch. Op de algoritmische problemen die opduiken bij het berekenen van die functies gaan we hier niet in. De parameterschattingen, zowel met CML als met MML, voor volledige en onvolledige designs zijn geïmplementeerd in het computerprogramma OPLM (Verhelst, Glas & Verstralen, 1993).

Voor de toetsing van het model kunnen we volstaan met een simpele verwijzing naar paragraaf 4.3: de rationale van de toetsen, maar ook hun technische uitwerking kan zonder meer toegepast worden op het meer algemene OPLM. Het is wel belangrijk, niet uit het oog te verliezen dat de vooraf gekozen discriminatie-indices deel uitmaken van het model en dus van de nulhypothese. Dit is analoog aan de situatie bij het LLTM, waar de gespecificeerde elementen van de $Q$-matrix eveneens deel uitmaken van de nulhypothese. De statistische toetsen hebben dus betrekking op het OPLM met de discriminatie-indices die door de gebruiker zijn gekozen. Een eventuele niet-passing van het model kan te wijten zijn aan de verkeerde specificatie van één of meer discriminatie-indices. De $S_i$-toetsen, maar vooral de $M_i$-toetsen kunnen gebruikt worden om dergelijke misspecificaties op het spoor te komen. De $M_i$-toetsen geven bovendien de richting aan waarin de discriminatie-index moet worden aangepast om een adequater model te krijgen. Werken met OPLM zal vaak bestaan uit het herhaaldelijk toepassen van de schattings- en toetsingsprocedures, waarbij iedere keer één of meer discriminatie-indices worden aangepast. Hoewel deze aanpassingen meestal gebeuren aan de hand van analyses op dezelfde data en er dus kanskapitalisatie kan optreden, is het belang van deze kanskapitalisatie gering als de steekproef niet te klein is. Meer beschouwingen hierover, alsook een heuristiek om plausibele waarden van de discriminatie-indices uit de data af te leiden, kan men vinden in Verhelst, Verstralen en Eggen (1991).

### *5.3.3 Modellen zonder de assumptie van lokale stochastische onafhankelijkheid*

Overtreding van het principe van de lokale stochastische onafhankelijkheid houdt in dat de onderlinge afhankelijkheid van itemantwoorden niet verdwijnt door te conditioneren op $\theta$. Dit betekent dat we kans op een antwoordpatroon gegeven $\theta$ niet kunnen schrijven als het produkt over items van de afzonderlijke kansen op een goed antwoord. Kelderman (1984, 1988) en Jannarone (1986) hebben een uitgebreide klasse van IRT-modellen beschreven waarin de kans op een antwoordpatroon rechtstreeks wordt gedefinieerd. We zien hier af van een complete beschrijving van deze klasse van modellen, omdat daarvoor een uitgebreid formalisme nodig is. In plaats daarvan zullen we het idee waarop een en ander gebaseerd is, toelichten aan de hand van een voorbeeld uit de klasse van modellen die door Jannarone is gedefinieerd. Stel dat een toets uit drie items bestaat. Beschouw een model waarin de kans op antwoordpatroon $\boldsymbol{x}$ gegeven $\theta$ geschreven kan worden als:

$$P(\boldsymbol{x} \mid \theta, \beta_1, \beta_2, \beta_3, \beta_{13}) = \frac{\exp\left(\sum_i x_i(\theta - \beta_i) + x_1 x_3(\theta - \beta_{13})\right)}{\sum_{\boldsymbol{y}} \exp\left(\sum_i y_i(\theta - \beta_i) + y_1 y_3(\theta - \beta_{13})\right)}, \tag{5.28}$$

waarbij het buitenste somteken in de noemer aangeeft dat de som genomen moet worden over alle mogelijke antwoordpatronen $\boldsymbol{y} = (y_1, y_2, y_3)$. In het voorbeeld heeft deze som dus acht termen. De functie van de noemer is er voor te zorgen dat de som van de kansen van alle acht antwoordpatronen gelijk is aan 1; voor de interpretatie is alleen de teller van belang. In dit model is er geen lokale stochastische onafhankelijkheid tussen de antwoordvariabelen $X_1$ en $X_3$. Dit kan formeel aangetoond worden door de formules voor $P(X_1 = 1 \mid \theta, X_3 = 1)$ en $P(X_1 = 1 \mid \theta, X_3 = 0)$ uit te schrijven zodat gedemonstreerd kan worden dat ze niet aan elkaar gelijk zijn. We kunnen echter de schending van de assumptie van lokale stochastische onafhankelijkheid ook duidelijk maken met een intuïtief argument. In de teller van (5.28) komen vier antwoordvariabelen aan bod: de drie itemantwoorden en het produkt $x_1 x_3$. Formeel kunnen we dit produkt opvatten als een vierde antwoord en dan is de teller van (5.28) niets anders dan de teller in de formule voor het Raschmodel met vier items. Doch er zijn slechts drie antwoorden geobserveerd en bijgevolg kunnen de vier itemantwoorden niet onafhankelijk zijn van elkaar. De noemer van (5.28) heeft dan ook geen 16 termen, want het produkt $y_1 y_3$ ligt volledig vast indien $y_1$ en $y_3$ gegeven zijn.

Merk op dat in dit model $\sum_i x_i + x_1 x_3$ de minimaal voldoende statistiek is voor $\theta$. Met andere woorden, als een respondent twee items juist heeft en zowel het eerste als het derde item is goed gemaakt, is de voldoende statistiek voor de vaardigheid groter dan wanneer het eerste en het tweede item goed worden gemaakt. Het simultaan goed maken van de items een en drie levert de persoon een extra scorepunt op voor de schatting van zijn vaardigheids- parameter. De parameter $\beta_{13}$ is de moeilijkheidsparameter die geassocieerd is met het be-halen van dit extra scorepunt.

Jannarone (1986) generaliseerde dit soort ideeën naar een zeer algemeen model. De parameters in dit model zijn te schatten met de CML-methode en er zijn toetsingsprocedures mogelijk die gebaseerd zijn op statistieken met een bekende asymptotische verdeling, in de lijn van de toetsingsprocedures die in hoofdstuk 4 zijn uiteengezet.

De modellen die door Kelderman (1984, 1988) zijn ontwikkeld, lijken erg veel op de modellen van Jannarone. Het essentiële verschil bestaat erin dat bij Kelderman de score gedefinieerd is als het aantal juiste itemantwoorden en niet meer afhangt van het produkt. In het voorgaande voorbeeld is de score 2 indien de persoon twee items juist heeft beantwoord, ongeacht welke twee dat zijn. Voor het voorbeeld (5.28) is de kans in Keldermans benadering gegeven door

$$P(\boldsymbol{x} \mid \theta, \beta_1, \beta_2, \beta_3, \beta_{13}) = \frac{\exp\left(\sum_i x_i(\theta - \beta_i) - x_1 x_3 \beta_{13}\right)}{\sum_{\boldsymbol{y}} \exp\left(\sum_i y_i(\theta - \beta_i) - y_1 y_3 \beta_{13}\right)}. \qquad (5.29)$$

Beide formules, (5.28) en (5.29), lijken erg op elkaar en het is ook niet zonder meer duidelijk wat de verschillen in interpretatie tussen beide benaderingen betekenen en of deze verschillen in de praktijk belangrijk zijn. De CML-procedure is in Keldermans benadering echter gemakkelijker toe te passen dan in Jannarones modellen, omdat de score onafhankelijk is van produkten van antwoordvariabelen. De klasse van modellen die Kelderman ontwikkelde is geïmplementeerd in het computerprogramma LOGIMO (Kelderman & Steen, 1988). De bestudering van Keldermans modellen is om nog een reden interessant. Kelderman bestudeerde het Raschmodel als een speciaal geval uit de klasse van de log-lineaire modellen en paste bij het schatten van de parameters ook technieken toe die veel gebruikt worden in de log-lineaire analyse.

Vooraleer we het laatste model uit deze paragraaf bespreken, moeten we nog even wat dieper ingaan op het begrip lokale stochastische onafhankelijkheid. In de definitie refereert het begrip 'lokaal' naar het feit dat er geconditioneerd wordt op de persoonsparameter $\theta$. Op het ogenblik dat de vaardigheid van de persoon verandert gedurende het maken van de toets, bijvoorbeeld ten gevolge van een leerproces of als gevolg van vermoeidheid of verveling is niet meer duidelijk op welke manier we nog van lokale stochastische onafhankelijkheid gebruik kunnen maken. Fischer (1972) heeft een benaderingswijze voor dit probleem bedacht die veel lijkt op de benadering met fysieke en conceptuele items die in paragraaf 5.1.3 werd gehanteerd. Stel dat er na het juist beantwoorden van een item een leerproces plaatsvindt, en dat de vaardigheid toeneemt met $\alpha$. Bij het beantwoorden van het zesde item beschikt persoon *v* dus over een vaardigheid $\theta_v + j\alpha$, waarin *j* het aantal correcte antwoorden is op de items 1 tot 5 en $\theta_v$ de vaardigheid bij het begin van de toetsafname. Maar dit is in de context van het Raschmodel hetzelfde als zeggen dat die persoon een vaardigheid $\theta_v$ heeft en dat het item een moeilijkheidsparameter heeft die gelijk is aan $\beta_6 + j\alpha$. We redeneren dus alsof we beschikken over zes conceptuele items in plaats van over één fysiek item. Elk conceptueel item correspondeert dus met een van de mogelijke waarden 0 tot en met 5 van *j*. Fischer heeft aangetoond dat met deze benadering geen CML-schattingen van de itemparameters en van de extra parameter $\alpha$ kunnen worden berekend waarna hij de hele benaderingswijze heeft opgegeven. Verhelst en Glas (1993) hebben echter aangetoond dat in het gegeven voorbeeld wel MML-schatters bestaan. Bovendien hebben zij aangetoond dat er andere situaties zijn waarin $\theta$ verandert gedurende de toetsafname, waar de CML-procedure wel kan worden toegepast.

We sluiten deze paragraaf af met een algemene beschouwing over het nut van de genoemde, misschien op het eerste gezicht nogal exotisch ogende modellen. De subtiele verschillen in interpretatie tussen de modellen van Kelderman en Jannarone kunnen de vraag doen rijzen of de vele inspanningen die onderzoekers zich getroosten om dergelijke, in het algemeen zeer ingewikkelde modellen te ontwikkelen enig praktisch nut hebben. Wij denken van wel en wel in om twee redenen.

Iedereen die enigszins bekend is met de wetenschappelijke psychologie, weet dat psychologische theorieën in elegantie en precisie niet kunnen wedijveren met bijvoorbeeld de theorieën in de natuurkunde. Een van de vele problemen waar de wetenschappelijke psychologie mee kampt, bestaat uit de vele op het eerste gezicht tegenstrijdige resultaten die in experimenten worden gevonden. De reden voor deze tegenstrijdigheden kan liggen in het gebrek aan precisie waarmee uitkomsten worden voorspeld, of in subtiele redeneringsfouten. Het construeren van formele modellen heeft het voordeel dat precieze predicties automatisch, dit wil zeggen langs wiskundige weg, uit een klein aantal veronderstellingen volgen. Het gevaar van subtiele fouten in de redenering is hierbij veel minder groot dan bij het gebruik van de natuurlijke taal.

Een tweede reden die voor de praktijk wellicht relevanter is, illustreren we met het volgende voorbeeld. Bij het construeren van examens is het in vele gevallen onvermijdelijk dat de items geformuleerd zijn als testlets, waarbij meer dan één vraag gesteld wordt bij dezelfde stam, bijvoorbeeld een inleidende tekst. De vragen worden meestal als aparte items beschouwd. Het is duidelijk dat het veel gemakkelijker is, lokale stochastische onafhankelijk- heid te realiseren tussen antwoorden op items die bij een verschillende stam behoren, dan tussen items die tot dezelfde stam horen. Het verkeerd lezen of interpreteren van de stam kan er de oorzaak van zijn dat alle items die bij die stam horen, verkeerd worden beantwoord. Daardoor is het principe van de lokale onafhankelijkheid geschonden en dat kan er de reden van zijn dat een eenvoudig IRT-model statistisch niet houdbaar is. Als men in zo'n geval toch het Raschmodel gebruikt en bijvoorbeeld de toetsscore definieert als het aantal items juist, betekent dit niet dat die scores 'waardeloos' zijn. Het kan wel betekenen dat iemand door één enkele onoplettendheid vier of vijf punten verliest, die anders wel behaald zouden zijn. Of iets algemener gezegd, de betrouwbaarheid van het resulterende meetinstrument, en dus ook de validiteit, zullen lager zijn dan wanneer een meetmodel werd gebruikt waarbij in deze afhankelijkheid werd voorzien, zoals de modellen van Jannarone en Kelderman. Vanuit deze optiek verschijnt het Raschmodel als een ideaaltype, waaraan in de praktijk vaak niet kan worden voldaan. De meer ingewikkelde modellen fungeren dan als een soort statistische correctieprocedure waarmee de vaak onvermijdelijke schendingen van het Raschmodel in de uiteindelijke meetresultaten kunnen worden

gecorrigeerd, analoog aan de manier waarop de covariantie-analyse gebruikt kan worden in quasi-experimenten, waar het ideaaltype van het gerandomiseerde experiment niet kan worden gerealiseerd.

## 5.4 Unidimensionale modellen voor polytome items

Dichotome items kunnen worden beschouwd als een speciaal geval van polytome items, waarbij het aantal antwoordcategorieën per item gelijk is aan twee. We kunnen dus ook het Raschmodel beschouwen als een speciaal geval van een model voor polytome items. Hoewel we in principe niets toe te voegen hebben aan de discussie over het Raschmodel die in hoofdstuk 4 is gevoerd, kunnen we bepaalde aspecten iets anders belichten, zodat de veralgemening naar modellen voor polytome items gemakkelijker wordt.

Het eerste aspect heeft te maken met het aantal responsfuncties per item dat nodig is om het model te definiëren. Omdat er twee antwoordcategorieën zijn, kunnen we in principe twee responsfuncties onderscheiden: de kans op een juist antwoord en de kans op een fout antwoord, beiden als functie van de latente variabele θ. Omdat de som van beide functies voor elke waarde van θ gelijk moet zijn aan 1, ligt de tweede functie volledig vast als de eerste gespecificeerd is. Er zijn dus wel twee functies maar er is slechts één onafhankelijke functie. Indien een item $m > 2$ antwoordcategeorieën heeft, kunnen we een responsfunctie beschouwen voor elk van de $m$ categorieën, maar de som van deze $m$ functies is de constante functie 1, zodat er slechts $m - 1$ onafhankelijke functies zijn. Deze functies dragen de naam categorieresponsfuncties. De itemresponsfunctie in het Raschmodel is dus de categorie- responsfunctie voor categorie 1.

Het tweede aspect betreft het aantal parameters per item. Men zou kunnen redeneren dat het natuurlijk is een parameter te associëren met elke categorie. Deze parameter zou dan als het ware de aantrekkingskracht uitdrukken die elke categorie uitoefent op de persoon die het item beantwoordt. Het is inderdaad mogelijk het Raschmodel op die manier op te schrijven:

$$P(X_i = 1|\theta) = \frac{\exp(1\,\theta - \eta_{i1})}{\exp(0\,\theta - \eta_{i0}) + \exp(1\,\theta - \eta_{i1})} = \frac{\exp(\theta - \eta_{i1})}{\exp(-\eta_{i0}) + \exp(\theta - \eta_{i1})}, \qquad (5.30)$$

waarin de coëfficiënten 1 en 0 van θ in het middelste lid van (5.30) het verschillende gewicht uitdrukken dat de twee antwoorden hebben met betrekking tot de latente

variabele θ. Het linkerlid van (5.30) blijft onveranderd indien in het rechterlid teller en noemer worden vermenigvuldigd met een constante die verschilt van nul. Kiezen we nu $\exp(\eta_{i0})$ als constante en definiëren we

$$\beta_i = \eta_{i1} - \eta_{i0}, \tag{5.31}$$

dan kunnen we (5.30) herschrijven als

$$P(X_i{=}1\,|\,\theta) = \frac{\exp[\theta - (\eta_{i1} - \eta_{i0})]}{1 + \exp[\theta - (\eta_{i1} - \eta_{i0})]} = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}. \tag{5.32}$$

De parameter $\beta_i$ kan dus geïnterpreteerd worden als het verschil tussen twee categorie-parameters. Deze parameters zelf zijn echter niet schatbaar.

Merk op dat de definitie van $\beta_i$ in (5.31) niet dwingend is. We hadden net zo goed teller en noemer van het rechterlid van (5.30) kunnen vermenigvuldigen met $\exp(\eta_{i1})$ en dit resulteert in

$$P(X_i{=}1\,|\,\theta) = \frac{\exp(\theta)}{\exp(\beta_i) + \exp(\theta)}, \tag{5.33}$$

maar dit is precies hetzelfde als (5.32).

Het derde aspect is impliciet reeds aan de orde gekomen in het middelste lid van (5.30), waar we de coëfficiënten van θ expliciet hebben opgeschreven. Een antwoord $X_i = 1$ resul-teert in een coëfficiënt 1 en een antwoord $X_i = 0$ heeft coëfficiënt 0. Dat wil zeggen dat de ordening van de coëfficiënten samenvalt met de ordening van de antwoordcategorieën en dat betekent dat de categorieën als geordende categorieën worden geïnterpreteerd. Het feit dat de coëfficiënten hier gelijk zijn aan de antwoorden is een extra eis die het Raschmodel aan de data oplegt. In het 2PL of OPLM is de ordening wel bewaard, doch de gelijkheid is opgegeven.

### *5.4.1 Het partial credit model (PCM)*

Gebruik makend van de drie voorgaande opmerkingen is de veralgemening van het Rasch-model tot een model voor polytome items voor de hand liggend. Het enige dat we moeten doen is nog een paar afspraken maken over de notatie. De categorieresponsfuncties zullen we aanduiden als $f_{ij}(\theta)$, waarbij de eerste index het item aanduidt en de tweede index de categorie. We hoeven daarbij niet aan te nemen dat elk item evenveel antwoordcategorieën heeft. Het aantal antwoordcategorieën per

item zullen we aanduiden als $m_i + 1$, waarbij de 'waarden' van de categorieën de opeenvolgende gehele getallen $0, 1, ..., m_i$ zijn. De veralgemening van (5.30) is dan gegeven door

$$f_{ij}(\theta) = P(X_i = j \mid \theta) = \frac{\exp(j\theta - \eta_{ij})}{\sum_{h=0}^{m_i} \exp(h\theta - \eta_{ih})}, \quad (j = 1, ..., m_i). \tag{5.34}$$

Voeren we nu de volgende herparametrisering in die analoog is aan (5.31):

$$\begin{aligned}
&\beta_{i0} = \eta_{i0} - \eta_{i0} = 0 \\
&\beta_{i1} = \eta_{i1} - \eta_{i0} \\
&\beta_{i2} = (\eta_{i2} - \eta_{i0}) - (\eta_{i1} - \eta_{i0}) = \eta_{i2} - \eta_{i1} \\
&. \\
&. \\
&\beta_{ij} = \eta_{ij} - \eta_{i,j-1} \\
&. \\
&. \\
&\beta_{i,m_i} = \eta_{i,m_i} - \eta_{i,m_i-1}
\end{aligned} \tag{5.35}$$

dan kan (5.34) geschreven worden als

$$f_{ij}(\theta) = \frac{\exp\left[j\theta - \sum_{g=0}^{j} \beta_{ig}\right]}{\sum_{h=0}^{m_i} \exp\left[h\theta - \sum_{g=0}^{h} \beta_{ig}\right]} = \frac{\exp\left[j\theta - \sum_{g=1}^{j} \beta_{ih}\right]}{1 + \sum_{h=1}^{m_i} \exp\left[h\theta - \sum_{g=1}^{h} \beta_{ig}\right]}, \tag{5.36}$$

waarin het rechterlid gelijk is aan het middelste lid omdat $\beta_{i0} = 0$. (De som-zonder-termen $\sum_{g=1}^{0} \beta_{ig}$ die voorkomt in geval $j = 0$, wordt daarbij gedefinieerd als 0.) Het model heeft dus maar $m_i$ vrije parameters per item want de parameterisering is zo gekozen dat $\beta_{i0} = 0$. Het model in zijn vorm (5.34) is voorgesteld door Andersen (1977), waarbij de achterliggende gedachte het ontwikkelen was van een veralgemening van het Raschmodel waarbij de score $s = \sum_i x_i$ een voldoende steekproefgrootheid voor $\theta$ is. De equivalente vorm (5.36) is door Masters (1982) voorgesteld onder de naam

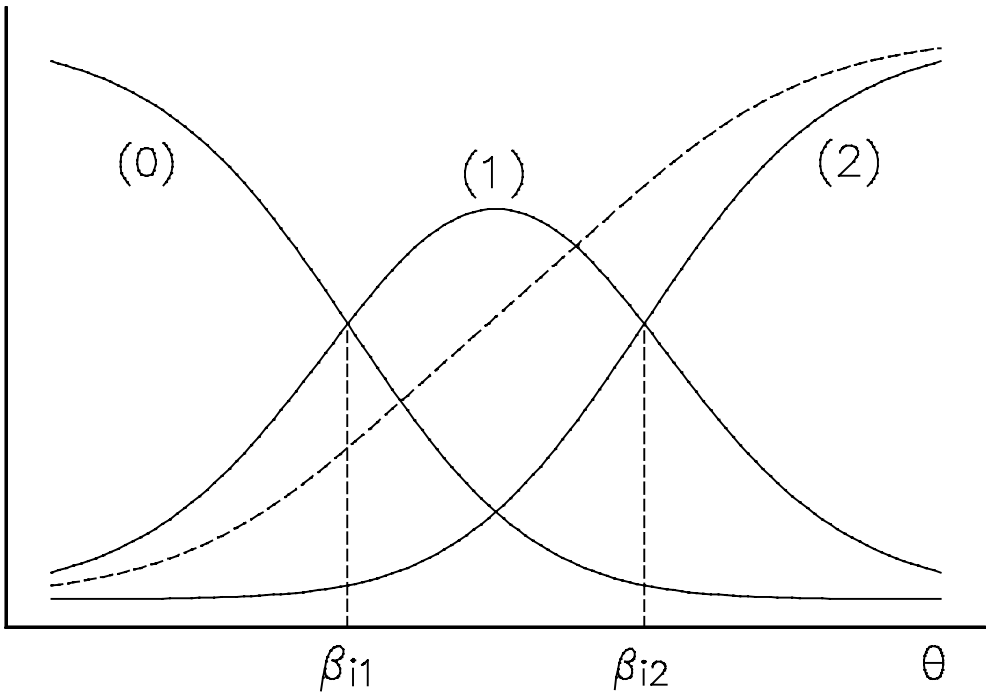
partial credit model (PCM). Om deze naam te begrijpen beschouwen we het volgende rekenitem dat ook door Masters werd gebruikt:

Bereken $\sqrt{7.5/0.3 - 16}$ .

Om dit item correct op te lossen moeten drie bewerkingen in de juiste volgorde correct worden uitgevoerd, een deling, een aftrekking en een worteltrekking. De achterliggende idee was om aan elke correct uitgevoerde stap een 'partial credit' toe te kennen. Men kon dus 0, 1, 2 of 3 punten verdienen bij de beantwoording van dit item. De idee van Masters was om voor elke stap op een of andere manier het Raschmodel te gebruiken. Indien we (5.36) gebruiken om de kans $P(X_i = j \mid \theta, X_i = j \text{ of } X_i = j - 1)$ te bepalen, dan krijgen we

$$P(X_i = j \mid \theta, X_i = j \text{ of } X_i = j - 1) = \frac{\exp(\theta - \beta_{ij})}{1 + \exp(\theta - \beta_{ij})}, \quad (j = 0,...,m_i). \tag{5.37}$$

Masters vertrok van (5.37) en toonde aan dat (5.36) daaruit volgt. Hoewel de benadering van Masters elegant is, dient men zich toch te hoeden voor twee conclusies die voor de hand lijken te liggen, maar die niet gerechtvaardigd zijn. De eerste betreft de betekenis van de parameters. Men zou kunnen denken dat in het voorgaande voorbeeld de parameter $\beta_{i2}$ de moeilijkheid aangeeft van de aftrekking 25-16. Deze conclusie is echter onjuist omdat de waarde van deze parameter ook beïnvloed wordt door de moeilijkheid van de daaropvolgende stap, de worteltrekking. In het algemeen kan men dus de parameters niet interpreteren als de moeilijkheid van de itemstappen. Molenaar (1983) heeft aan dit probleem een uitvoerige discussie gewijd. Een tweede misvatting ontstaat indien men denkt dat het PCM alleen geldig kan zijn bij items die in stapjes kunnen worden onderverdeeld. In feite treedt hier hetzelfde probleem op als we besproken hebben bij het 3PL. De stapjesrationale van Masters is een cognitief model dat tot het PCM leidt, maar het omgekeerde volgt niet noodzakelijk, net zo min als uit het 3PL het cognitief model volgt dat in paragraaf 5.3.1 werd besproken. Voor een voorbeeld waar de stapjesidee zeker niet van toepassing is, maar het PCM wel, zie Verhelst en Verstralen (1991). De interpretatie van de categorieparameters kunnen we het beste begrijpen aan de hand van figuur 5.5 waar de categorieresponsfuncties en de itemregressiefunctie zijn getekend voor een item $i$ met $m_i = 2$. De categorieën zijn tussen haakjes aangeduid in de figuur.



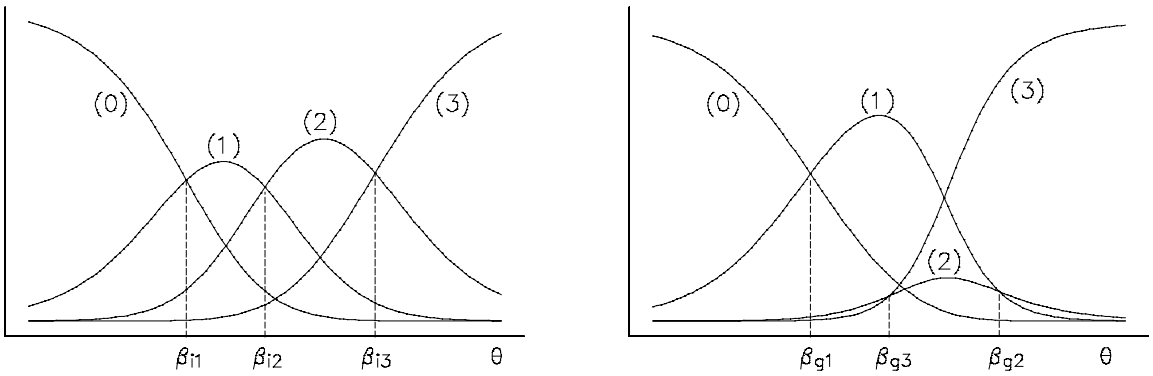

Figuur 5.5
Categorieresponsfuncties voor een item met drie antwoordcategorieën

De parameter $\beta_{i1}$ geeft aan waar de responscurven voor categorie 1 en 0 elkaar snijden en de parameter $\beta_{i2}$ komt overeen met het snijpunt van de categorieën 1 en 2. In het algemeen is de parameter $\beta_{ij}$ die waarde van de latente variabele θ waarvoor de categorieën $j$ en $j-1$ een even grote kans hebben gekozen te worden. Merk op dat dit ook geldt in het Raschmodel. De itemparameter $\beta_i$ kunnen we ook interpreteren als de categorieparameter $\beta_{i1}$, dus als die waarde van θ waar beide categorieën een even grote kans hebben. Omdat er slechts twee categorieën zijn, is die kans gelijk aan 0.5.

De curve in stippellijnen in figuur 5.5 is een kleine modificatie van de itemregressiefunctie. Het is de curve van de functie $\mathscr{E}(X_i \mid \theta)/m_i$, die men de gestandaardiseerde itemregressie-functie kan noemen. De categorieresponscurve voor de middelste categorie is eentoppig. In het algemeen geldt in het PCM dat de curve voor categorie 0 monotoon dalend is in θ, de curve voor categorie $m_i$ is monotoon stijgend en alle andere curven zijn eentoppig. De item-regressiefunctie echter is monotoon stijgend en dat is de reden waarom we items die aan het PCM voldoen monotone items noemen.

In figuur 5.5 is duidelijk dat categorie 1 de grootste kans heeft als θ in het interval $(\beta_{i1}, \beta_{i2})$ ligt. De uitspraak 'categorie $j$ $(j = 1,..., m_i - 1)$ is de modale categorie in het interval $(\beta_{ij}, \beta_{i,j+1})$' is slechts juist indien men beseft dat dit interval alleen bestaat

indien $\beta_{ij} < \beta_{i,j+1}$ en dat deze ongelijkheid niet door het model verondersteld wordt. In figuur 5.6 zijn de categorieresponscurven afgebeeld voor twee items *i* en *g*. Voor item *i* geldt dat $\beta_{i1} < \beta_{i2} < \beta_{i3}$, maar voor item *g* geldt dat $\beta_{g2} > \beta_{g3}$.




Figuur 5.6
Geordende en niet-geordende categorieparameters

Voor item *i* geldt voor alle categorieën dat ze modaal, dat is het waarschijnlijkst, zijn in een bepaald interval van θ. Voor item *g* geldt dit niet, want categorie 2 is nooit de meest waarschijnlijke categorie. Merk op dat de waarden van θ waarvoor de categorierespons- functies van de verschillende categorieën hun grootste waarde bereiken wel degelijk geordend zijn in dezelfde volgorde als de categorieën. Zo geldt voor beide items in figuur 5.6 dat de θ-waarde waar categorie 2 haar grootste kans bereikt, groter is dan de θ-waarde waar categorie 1 haar grootste kans bereikt.

Het schatten van de parameters in het PCM kan met CML of MML gebeuren. Om de schattingsvergelijkingen op een elegante manier te kunnen opschrijven, voeren we een indicatorvector $\boldsymbol{Y}_{vi}$ in die $m_i$ elementen bevat. Indien de antwoordvariabele $X_{vi}$ gelijk is aan 0, zijn alle $m_i$ elementen van $\boldsymbol{Y}_{vi}$ eveneens gelijk aan 0. Indien $X_{vi} = j$, dan is het *j*-de element van $\boldsymbol{Y}_{vi}$ gelijk aan 1, de andere elementen zijn gelijk aan 0. De vectoren $\boldsymbol{Y}_{vi}$ bevatten dus precies dezelfde informatie als de oorspronkelijke antwoordvariabelen. De elementen van de vector $\boldsymbol{Y}_{vi}$ zullen we in het algemeen aanduiden als $Y_{vij}$. Bijvoorbeeld, indien $m_i = 4$, dan geldt

$$X_{vi} = 3 \Leftrightarrow \boldsymbol{Y}_{vi} = (0, 0, 1, 0).$$

De geobserveerde antwoorden van persoon *v* kunnen we dus schrijven als één lange vector $\boldsymbol{Y}_v$ door alle vectoren $\boldsymbol{Y}_{vi}, (i = 1,...,k)$ gewoon achter elkaar te schrijven. De matrix *Y* van observaties krijgen we dan door de *n* vectoren $\boldsymbol{Y}_v$ in een tabel onder

elkaar te schrijven. Door gebruik te maken van het axioma van de lokale stochastische onafhankelijkheid kan de log-aannemelijkheidsfunctie gegeven één enkele vector $\boldsymbol{Y}_v$ geschreven worden als

$$\ln L(\theta_v, \beta; \boldsymbol{y}_v) = s_v\theta_v + \sum_{j=1}^{m_i} y_{vij}\left(-\sum_{g=1}^{j}\beta_{ig}\right) - \sum_{i=1}^{k}\ln\left(1 + \sum_{h=1}^{m_i}\exp[h\theta_v - \sum_{g=1}^{h}\beta_{ig}]\right), \qquad (5.38)$$

waarin

$$s_v = \sum_{i}^{k}\sum_{j}^{m_i} j\, y_{vij} = \sum_{i}^{k} x_{vi}$$

de score is van persoon *v*, dat wil zeggen het totaal aantal 'punten' dat persoon *v* behaald heeft. Definiëren we nu

$$t_{ij} = \sum_{v} y_{vij},$$

en maken we gebruik van (5.35), dan kan de log-aannemelijkheidsfunctie gegeven de antwoorden van *n* geschreven worden als

$$\ln L(\theta, \beta; Y) = \sum_{v} s_v\theta_v + \sum_{j=1}^{m_i} t_{ij}(-\eta_{ij}) - \sum_{v}\sum_{i=1}^{k}\ln\left(1 + \sum_{h=1}^{m_i}\exp(h\theta_v - \eta_{ih})\right). \qquad (5.39)$$

Het is duidelijk dat (5.39) een log-aannemelijkheidsfunctie is uit de exponentiële familie en dat bovendien kan geconditioneerd worden op de voldoende steekproefgrootheid voor $\theta_v$. Op analoge wijze als bij het Raschmodel en bij het OPLM voor dichotome data kan de conditionele log-aannemelijkheidsfunctie geschreven worden als

$$\ln L(\boldsymbol{\varepsilon}; X \mid \boldsymbol{s}) = \sum_{i}^{k}\sum_{j}^{m_i} t_{ij}\ln\varepsilon_{ij} - \sum_{v}\ln\gamma_{s_v}(\boldsymbol{\varepsilon}), \qquad (5.40)$$

waarin

$$\varepsilon_{ij} = \exp(-\eta_{ij}) = \exp(-\sum_{g=1}^{j}\beta_{ij})$$

en

$$\gamma_s(\boldsymbol{\varepsilon}) = \sum_{\Sigma_i x_i = s}\varepsilon_{ij}^{y_{ij}}. \qquad (5.41)$$

De functie $\gamma_s(\boldsymbol{\varepsilon})$ is een veralgemening van de symmetrische basisfuncties die in het Rasch- model werden gebruikt. Het rechterlid van (5.41) geeft aan dat de som genomen moet worden over alle antwoordpatronen die de score *s* opleveren. De analogie met het Raschmodel komt verder tot uiting in de conditionele schattingsvergelijkingen die we hier zonder gedetailleerde afleiding weergeven:

$$t_{ij} = \sum_v \pi_{ij|s_v} = \sum_v \frac{\varepsilon_{ij}\gamma^{(i)}_{s_v-j}(\boldsymbol{\varepsilon})}{\gamma_{s_v}(\boldsymbol{\varepsilon})}, \tag{5.42}$$

waarin $\tau_{ij|s}$ een verkorte notatie is van $P(X_i = j \mid s)$. Het superscript (*i*) bij het functiesymbool $\gamma$ geeft aan dat alle categorieparameters $\varepsilon_{ij}, (j = 1,...,m_i)$ uit de argumentvector $\boldsymbol{\varepsilon}$ moeten worden weggelaten.

De schattingsvergelijkingen voor MML zijn eveneens in analogie met het Raschmodel op te stellen. We gaan er hier niet nader op in. Zowel CML-schattingen als MML-schattingen voor de parameters in het PCM kunnen met het computerprogramma OPLM worden berekend. De statistische toetsing van het PCM wordt in de volgende paragraaf besproken.

### *5.4.2 Generalisaties van het partial credit model*

#### *OPLM voor polytome items*

Hoewel we gezien hebben dat in het PCM het aantal categorieën per item verschillend mag zijn, levert het hanteren van verschillende aantallen bij het construeren van een toets soms moeilijkheden op. Veronderstel dat een toetsconstructeur over twee items beschikt die hij graag in eenzelfde toets wil opnemen. Het eerste item leent zich uitstekend om partieel gescoord te worden, waarbij de constructeur duidelijke voorschriften heeft wanneer een antwoord 0, 1 of 2 punten verdient. Voor het andere item ligt deze partiële scoring echter niet voor de hand, zodat alleen dichotome scoring overblijft. Binnen het PCM levert een correct antwoord op het eerste item 2 punten op, terwijl een correct antwoord op het tweede item slechts 1 punt oplevert. De twee items worden dus verschillend gewogen en deze weging volgt automatisch uit het aantal antwoordcategorieën. Dergelijke automatische koppeling kan zeer contra-intuïtief zijn en een reden waarom het PCM slechte passing geeft indien er grote variabiliteit is in het aantal antwoordcategorieën per item. Een veralgemening van het model die aan dit bezwaar tegemoetkomt ontstaat door het toevoegen van een verschillend gewicht per

item. Dit gewicht duiden we aan als $a_i$. De itemresponsfunctie voor deze veralgemening van het PCM is gegeven door een eenvoudige verandering van (5.34):

$$f_{ij}(\theta) = P(X_i = j \mid \theta) = \frac{\exp[a_i(j\theta - \eta_{ij})]}{\sum_{h=0}^{m_i} \exp[a_i(h\theta - \eta_{ih})]}, \quad (j = 1, ..., m_i). \tag{5.43}$$

Afhankelijk van de status die men aan de grootheid $a_i$ toekent ontstaan polytome generalisaties van twee modellen die we reeds eerder hebben besproken. Beschouwen we de grootheden $a_i$ als onbekende parameters die uit de data moeten worden geschat, dan is (5.43) een veralgemening van het 2PL, beschouwen we ze echter als gekende indices, dan krijgen we een polytome veralgemening van het OPLM. Willen we, zoals in het voorbeeld hierboven, alle items even zwaar laten wegen, ongeacht het aantal antwoordcategorieën, dan krijgen we een speciaal geval van het OPLM waarbij de $a_i$ proportioneel zijn met $1/m_i$. De generalisatie (5.43) waarbij de $a_i$ behandeld worden als te schatten parameters is in de literatuur niet beschreven als een unidimensionaal model. In paragraaf 5.5 zullen we echter zien dat het weer opduikt als een speciaal geval van een multidimensionaal model.

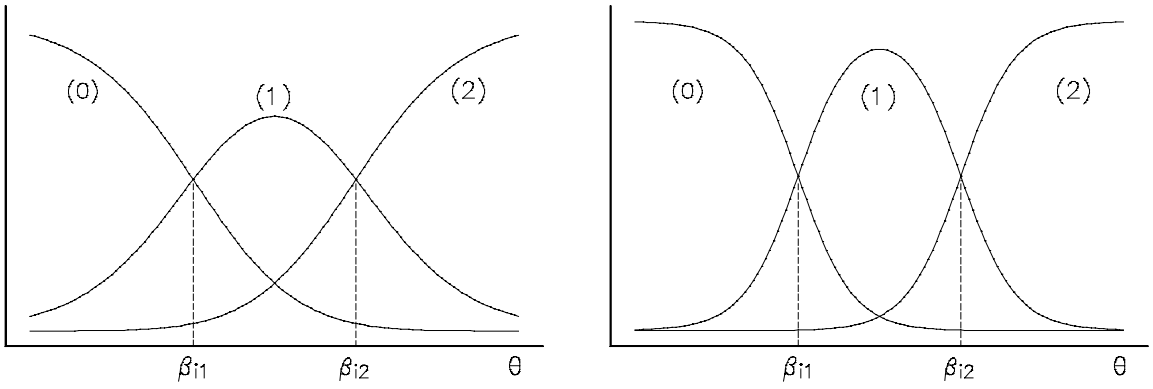
De generalisatie (5.43) waarbij de $a_i$ bekende constanten zijn, die bovendien alleen gehele waarden aannemen, zullen we verder kortheidshalve aanduiden als het polytome OPLM. Schattingen van de parameters, zowel met CML als met MML, kunnen met het computer- programma OPLM berekend worden. Voor technische details verwijzen we naar Verhelst, Glas en Verstralen (1993).

De statistische toetsen voor het polytome OPLM en dus ook voor het PCM, zijn veralgemeningen van de statistische toetsen voor het Raschmodel en spreken meestal voor zich. Zo is bijvoorbeeld de benaderende kwadratische vorm $R_{1c}^*$ die in (4.101) werd gegeven in de context van het Raschmodel, in het geval van het polytome OPLM gegeven door

$$R_{1c}^* = \sum_{q=1}^{r} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\left[\sum_{s \in G_q} n_s(p_{ij|s} - \hat{\pi}_{ij|s})\right]^2}{\sum_{s \in G_q} n_s \hat{\pi}_{ij|s}(1 - \hat{\pi}_{ij|s})}, \tag{5.44}$$

waarin de scores worden opgedeeld in $r$ scoregroepen $G_q$, $(q = 1,...,r)$. Voor de $M_i$- en de $S_i$-toetsen treedt echter een complicatie op, die onmiddellijk duidelijk wordt indien

we figuur 4.7 bekijken vanuit het standpunt van modelpassing bij polytome items. De voorspelde waarden in die figuur hebben betrekking op categorie 1 van het item *i* en een systematische onder- of overschatting van de discriminatie-index wordt onmiddellijk duidelijk uit een steiler respectievelijk vlakker verloop van de geobserveerde proporties in vergelijking met de voorspelde proporties. Deze duidelijkheid gaat echter verloren indien we analoge figuren construeren voor de middencategorieën bij polytome items. Dit is goed te zien in figuur 5.7.




Figuur 5.7
Responscurven voor een polytoom item met $a_i = 1$ (links) en $a_i = 2$ (rechts)

In de figuur rechts is de discriminatie-index twee keer zo groot als in de figuur links. Stel nu dat $a_i$ in werkelijkheid gelijk is aan 1, doch we hebben ten onrechte gesteld dat $a_i = 2$. Als we nu, analoog aan figuur 4.7 een curve construeren waarin we $\hat{\pi}_{i1|s}$ en $p_{i1|s}$ uitzetten tegen de score *s*, dan zullen voorspelde proporties ongeveer het patroon volgen van de eentoppige curve rechts in figuur 5.7 en de geobserveerde proporties zullen het patroon volgen van de middelste curve uit het linkergedeelte van figuur 5.7. Deze beschrijving is echter nog een beetje geflatteerd omdat bij verkeerde specificatie van de discriminatie-indices ook de categorieparameters systematisch verkeerd geschat worden. Kortom, afwijkingen tussen voor- spelde en geobserveerde proporties bij de middencategorieën zijn wel systematisch, doch het is helemaal niet duidelijk hoe de scores moeten gegroepeerd worden om de statistische toetsen onderscheidend vermogen te geven tegen de verkeerde specificatie van de discriminatie-indices. In het programma OPLM is een oplossing gevonden voor dit probleem door de items na de schatting te dichotomiseren. Dichotomiseren we een item met 3 antwoordcategorieën door het antwoord 0 als lage categorie te beschouwen en de antwoorden 1 en 2 als hoge categorie, dan kunnen we voor de toetsing dezelfde rationale volgen als bij dichotome items. Definiëren we nu meer in het algemeen

$$\hat{\pi}^{*}_{ij|s} = \sum_{g=j}^{m_i} \hat{\pi}_{ig|s},$$

$$p^{*}_{ij|s} = \sum_{g=j}^{m_i} p_{ig|s},$$

dan is de veralgemening van de benaderende vorm $S_i^*$ (formule 4.98) voor het polytome geval gegeven door

$$S^{*}_{ij} = \sum_{q=1}^{r} \frac{\left[\sum_{s\in G_q} n_s(p^{*}_{ij|s} - \hat{\pi}^{*}_{ij|s})\right]^2}{\sum_{s\in G_q} n_s \hat{\pi}^{*}_{ij|s}(1 - \hat{\pi}^{*}_{ij|s})}, \qquad (j = 1,...,m_i). \tag{5.45}$$

Per item zijn dus $m_i$ toetsen beschikbaar, één voor elke dichotomisering van het item. Dichotomisering kan ook worden toegepast voor de $M_i$-toetsen. Voor toepassingen van deze toetsen zij men verwezen naar hoofdstuk 7 en hoofdstuk 9.

Terzijde kan nog worden opgemerkt dat de formules (5.44) en (5.45) geen rekening houden met de covariantie tussen de schatters van de categorieparameters. Bij parameters die tot het zelfde item behoren is de covariantie in absolute waarde heel wat groter dan bij parameters die tot verschillende items behoren. In de benaderende vormen van de toetsingsgrootheden die door het programma OPLM worden berekend, wordt alleen die laatste covariantie verwaarloosd; met de eerste wordt wel rekening gehouden. De formules worden hier niet gegeven omdat ze niet louter met sommen kunnen uitgedrukt worden.

### *De uitbreiding van het PCM door Wilson en Masters*

De schattingsvergelijkingen (5.42) in het PCM hebben niet altijd een oplossing. Een noodzakelijke voorwaarde is dat elke categorie, inclusief de nulcategorie, van elk item in de steekproef minstens één maal geobserveerd is. Indien een categorie in de steekproef niet geobserveerd is, dan gaan Wilson en Masters (1993) het model een beetje aanpassen, om de andere parameters toch te kunnen schatten. Stel dat met item *i* bij de constructie een scoringsregel is opgesteld die resulteert in vijf geordende categorieën van 0 tot 4, doch dat in de steekproef categorie 2 niet wordt geobserveerd. Het item wordt dan omgevormd tot een item met vier antwoordcategorieën, die

respectievelijk gewicht of score 0, 1, 3 en 4 krijgen. Om te zien hoe dit probleem opgelost kan worden, herschrijven we (5.43) in een iets gewijzigde vorm:

$$f_{ij}(\theta) = \frac{\exp(j a_i \theta - a_i \eta_{ij})}{\sum_{h=0}^{m_i} \exp(h a_i \theta - a_i \eta_{ih})} = \frac{\exp(A_{ij}\theta - \delta_{ij})}{\sum_{h=0}^{m_i} (A_{ih}\theta - \delta_{ih})}. \qquad (5.46)$$

Het rechterlid van (5.46) kunnen we beschouwen als een generieke gedaante van veel unidimensionale modellen voor polytome items. We zien dat de grootheid $a_i$ opgeslorpt is in de nieuwe categorieparameter $\delta_{ij}$, doch dit is geen probleem want door een simpele deling krijgen we de oorspronkelijke $\eta$-parameters terug. De verschillende modellen onderscheiden zich vooral van elkaar door de structuur en de status van $A_{ij}$, het gewicht of de score die aan een antwoord in de *j*-de categorie op item *i* moet worden toegekend. Zo kunnen we zeggen dat de categorieresponsfuncties van het PCM gegeven zijn door het rechterlid van (5.46), met $A_{ij} = j$. In tabel 5.2 wordt een overzicht gegeven van alle unidimensionale modellen die in dit boek behandeld worden als speciale gevallen van de algemene gedaante (5.46). De enige uitzondering is het 3PL, dat niet in deze categorisering past.

Tabel 5.2
Unidimensionale modellen als speciaal geval van (5.46)

| Model | $A_{ij}$ | Opmerkingen |
|---|---|---|
| Raschmodel | 0 en a | 0 voor een fout antwoord;<br>$a > 0$ voor een juist antwoord. |
| Dichotome OPLM | 0 en $a_i$ | 0 voor een fout antwoord;<br>$a_i$ een positief geheel getal voor een juist antwoord;<br>$a_i$ a priori vastgelegd. |
| 2PL | 0 en $a_i$ | 0 voor een fout antwoord;<br>$a_i > 0$, uit de data geschat. |
| PCM | $j$ | $j = 0,...,m_i$ |
| Polytome OPLM | $ja_i$ | $j = 0,...,m_i$;<br>$a_i$ is een positief geheel getal, a priori vastgelegd. |
| Polytome 2PL | $ja_i$ | $j = 0,...,m_i$;<br>$a_i > 0$, uit de data geschat. |

| | | |
|---|---|---|
| Wilson en Masters | $\ell_j$ | $\ell_j$ is een positief geheel getal a priori vastgelegd (alleen voor geobserveerde categorieën). |
| nominale responsmodel | $a_{ij}$ | uit de data geschat. |

De uitbreiding van het PCM die Wilson en Masters behandelen, kan ook als een speciaal geval (5.46) beschreven worden: zij kiezen voor $A_{ij}$ van te voren, door de scoringsregel, vastgelegde gehele waarden. In het voorbeeld dat we hierboven gaven geven zij voor de vier geobserveerde categorieën respectievelijk de gewichten 0, 1, 3 en 4.

We hebben reeds eerder gezien dat het model dat door (5.46) gegeven is, niet identificeerbaar is. Als een item 5 antwoordcategorieën heeft, dan verschijnen in (5.46) ook 5 categorieparameters, $\eta$ of $\delta$, voor dat item, doch ze zijn niet allemaal schatbaar. We hebben dit probleem opgelost door in het middelste lid van (5.46) teller en noemer te vermenigvuldigen met $\exp(\eta_{i0})$ en het spreekt vanzelf dat we dezelfde techniek kunnen toepassen op het rechterlid van (5.46) door teller en noemer te vermenigvuldigen met $\exp(\delta_{i0})$. In het bovenstaande voorbeeld heeft item *i* dus vijf categorieparameters, waarbij in de toepassing van Wilson en Masters er slechts drie geschat worden. De parameter $\delta_{i2}$ wordt niet geschat omdat de tweede categorie niet geobserveerd is en de drie overige parameters die wel geschat worden zijn de verschillen $\delta_{i1}-\delta_{i0}, \delta_{i3}-\delta_{i0}$ en $\delta_{i4}-\delta_{i0}$. Het is belangrijk hierbij op te merken dat de $\delta$-parameter die 'weggewerkt' wordt om het model identificeerbaar te maken, hier dus $\delta_{i0}$, niet mag overeenkomen met een categorie die niet geobserveerd is. Indien categorie 0 in de steekproef niet geobserveerd is kan $\exp(\delta_{i0})$ als factor in teller en noemer in het rechterlid van (5.46) om het model te identificeren. Doch zoals we reeds eerder zagen kan een willekeurige andere parameter, waarvan de overeenkomende categorie wel is geobserveerd, gebruikt worden. Dit maakt de interpretatie van de parameters er echter niet gemakkelijker op.

Hoewel de benadering van Wilson en Masters elegant is om parameters van polytome items te schatten indien niet alle categorieën geobserveerd zijn, moet het praktische nut van hun methode niet overschat worden. Indien in de calibratiesteekproef een bepaalde categorie niet voorkomt, dan heeft men geen schatting van de bijbehorende categorieparameter. Doch dit sluit niet uit dat bij een latere toepassing die categorie wel wordt geobserveerd. Dan is het niet mogelijk uit een antwoordpatroon waar deze categorie in voorkomt $\theta$ te schatten, omdat voor een schatting van $\theta$ de ontbrekende waarde van de categorieparameter nodig is.

### *Het nominale responsmodel*

Het rechterlid van (5.46) suggereert een verdere uitbreiding van het PCM. We kunnen namelijk het standpunt innemen dat we helemaal niets weten over de gewichten $A_{ij}$ en ze behandelen als parameters die uit de data moeten geschat worden. Doch dit impliceert dat $A_{i,j+1}$ kleiner kan zijn dan $A_{i,j}$, dus dat een antwoord in categorie $j$ hoger moet gewaardeerd worden dan een antwoord in categorie $j+1$. De ordening van de categorieën komt niet meer overeen met de ordening van hun gewichten. De categorienummers zijn dus gewoon labels van de categorie geworden en het resulterend model wordt dan ook het nominale responsmodel genoemd. Het werd voorgesteld door Bock (1972).

Het is niet moeilijk om uit het rechterlid van (5.46) af te leiden dat de voldoende steekproefgrootheid voor θ gegeven is door

$$\Sigma_i \Sigma_j A_{ij} y_{vij}. \tag{5.47}$$

Indien de gewichten $A_{ij}$ a priori zijn vastgelegd zoals in het PCM, het polytome OPLM en het model van Wilson en Masters, is deze grootheid zonder meer uit de data te berekenen en kan er dus op geconditioneerd worden. In deze modellen is CML dus mogelijk. In het nominaal respons model moeten de gewichten $A_{ij}$ geschat worden en kunnen dus niet gebruikt worden om te conditioneren. De MML-schattingsprocedure is wel mogelijk en is geïmple-menteerd in het computerprogramma MULTILOG (Thissen, 1988).

### *Het rating scale model*

In paragraaf 5.1 hebben we gezien dat het LLTM een specificatie is van het Raschmodel die ontstaat door op de itemparameters lineaire restricties op te leggen. Dit is natuurlijk ook mogelijk bij polytome items; alleen dient men een zinvolle theorie of hypothese voor deze restricties te hebben of te construeren. We bespreken hier één voorbeeld van dergelijke restricties, het rating scale model van Andrich (1978a, 1978b).

Een rating scale is een observatie-instrument waarbij een persoon uit een aantal geordende categorieën er een uitkiest die het beste zijn mening weerspiegelt met betrekking tot een bepaalde uitspraak of een bepaald onderwerp. We geven twee voorbeelden van items die van deze techniek gebruik maken.

Item A: Den Uyl was een goede premier van Nederland.
sterk oneens oneens eens sterk eens

Item B: De colleges van prof. P. zijn interessant.
sterk oneens oneens eens sterk eens

Item A is bedoeld om de politieke attitude te meten van de persoon die het item beantwoordt en item B wordt gebruikt in een vragenlijst die bedoeld is om de attitude ten opzichte van een bepaalde onderwijsinstelling te meten. Hoewel het formaat van beide items identiek is en beide items bedoeld zijn om een attitude te meten, volgt daar niet uit dat het gedrag met betrekking tot beide items met eenzelfde soort model adequaat kan worden beschreven. Als we, net als in paragraaf 5.2, de politieke attitude interpreteren als de traditionele 'links-rechts' dimensie, ligt het voor de hand item A te interpreteren als een niet monotoon item. Personen met een ultra-linkse of ultra-rechtse overtuiging zullen het waarschijnlijk met de uitspraak in item A niet eens zijn, hoewel ze op de veronderstelde dimensie zeer ver van elkaar gelokaliseerd zijn. Voor dit item lijkt het dus redelijk een model voor niet-monotone items te gebruiken. Bij item B daarentegen lijkt het redelijk aan te nemen dat personen die het zelfde antwoord geven niet drastisch van elkaar verschillen in hun attitude. Bovendien lijkt het redelijk aan te nemen dat de categorie 'sterk eens' wijst op een positievere attitude dan de categorie 'eens' of 'oneens'. Kortom, de interpretatie van item B als een monotoon item is veel aannemelijker dan dit het geval is bij item A. Het rating scale model van Andrich is ontwikkeld als model voor items die geïnterpreteerd worden als monotone items.

Het is kenmerkend voor het gebruik van rating scales dat de antwoordcategorieën waaruit gekozen moet worden allemaal op dezelfde manier gelabeld zijn. In het model van Andrich is de kans dat een persoon *v* op item *i* met categorie *j* antwoordt, afhankelijk van de latente attitude $\theta_v$ van die persoon, van de 'moeilijkheid' van het item *i* en van de 'moeilijkheid' van antwoordcategorie *j*. Om een goed begrip te hebben van het onderscheid tussen beide moeilijkheden beschouwen we nog een ander item uit de schoolattitudevragenlijst:

Item C: Prof. P. is de ideale lesgever.
sterk oneens oneens eens sterk eens

Een persoon die het sterk eens is met de uitspraak in item B hoeft het niet sterk eens te zijn met de uitspraak in item C. Met andere woorden item C is 'moeilijker' dan item

B. We hadden natuurlijk ook een vragenlijst kunnen construeren waarin we dezelfde uitspraken gebruikten als in de items B en C, maar de antwoordcategorieën formuleerden als: 'nee' en 'ja'. Het zal wel duidelijk zijn dat er een positievere attitude vereist is om het antwoord 'sterk eens' te kiezen dan het veel minder sterk gekleurde antwoord 'ja'. De categorie 'ja' impliceert een lagere drempel dan de categorie 'sterk eens'.

Het rating scale model van Andrich is een speciaal geval van het PCM waar de categorieparameter $\beta_{ij}$ uit formule (5.36) geschreven wordt als

$$\beta_{ij} = \gamma_i + \tau_j, \quad (i = 1,...,k\,;\, j = 1,...,m), \tag{5.48}$$

waarin $m$ het gemeenschappelijke aantal antwoordcategorieën is, $\gamma_i$ de itemparameter van item $i$ en $\tau_j$ de parameter van antwoordcategorie $j$. De parameters $\gamma$ en $\tau$ kunnen we dus opvatten als basisparameters; de categorie-parameters $\beta_{ij}$ van het PCM zijn dus lineaire combinaties van de basisparameters.

Naast het rating scale model van Andrich bestaan er nog andere interessante modellen, die kunnen geschreven worden als restricties op de PCM-parameters $\beta_{ij}$, doch in die gevallen gaat het niet meer om lineaire restricties. Details over deze modellen kan men vinden in Masters en Wright (1984).

## 5.5 Multidimensionale IRT-modellen

Het begrip unidimensionaliteit dat tot hier toe is gehanteerd, is redelijk eenduidig; het begrip multidimensionaliteit heeft vele betekenissen. Vooraleer we specifieke modellen aan de orde stellen, geven we een overzicht van de verschillende betekenissen van het begrip.

Grosso modo kunnen we twee klassen van multidimensionale benaderingen binnen de IRT onderscheiden. De eerste klasse betreft modellen die een beperkt probleem oplossen. De verzameling items die moet worden geanalyseerd is reeds opgedeeld in een aantal groepen items en voor elk van die groepen weet of veronderstelt men dat ze geschaald kunnen worden met een unidimensionaal IRT-model, bijvoorbeeld met het Raschmodel. Bij de tweede klasse van modellen weet men dit niet, of wenst men die veronderstelling niet te maken. Modellen die tot die klasse behoren zijn bedoeld om de multidimensionale structuur van de items te ontrafelen. Deze vage noties worden nu explicieter gemaakt.

Veronderstel dat men de beschikking heeft over een aantal toetsen, zeg $Q$, die elk adequaat beschreven kunnen worden door een unidimensionaal IRT-model. Elk van deze toetsen is dus geschikt om een latente eigenschap $\theta_q, (q = 1,..., Q)$, te meten. De vraag die men zich kan stellen is of deze $Q$ eigenschappen iets met elkaar te maken hebben, hoe groot bijvoorbeeld de correlatie tussen die eigenschappen is in een bepaalde populatie. Een voorbeeld van deze benadering wordt besproken in paragraaf 5.5.1.

In de tweede klasse van modellen wordt er van uitgegaan dat elk item een beroep doet op twee of meer latente vaardigheden. Deze modellen zijn bedoeld om na te gaan in welke mate elk item uit een toets een beroep doet op elke vaardigheid. Een mogelijke situatie is dat een gedeelte van de items uitsluitend een beroep doet op één vaardigheid en de overige items uitsluitend een andere vaardigheid aanspreken. Het zou echter ook kunnen zijn dat alle items op alle vaardigheden in verschillende aanspreken. Het is echter niet zonder meer duidelijk wat bedoeld wordt met uitdrukkingen als: ’een beroep doen op’ of ’aanspreken’. Deze begrippen dekken een heel complexe lading, die we met enkele voorbeelden zullen toelichten.

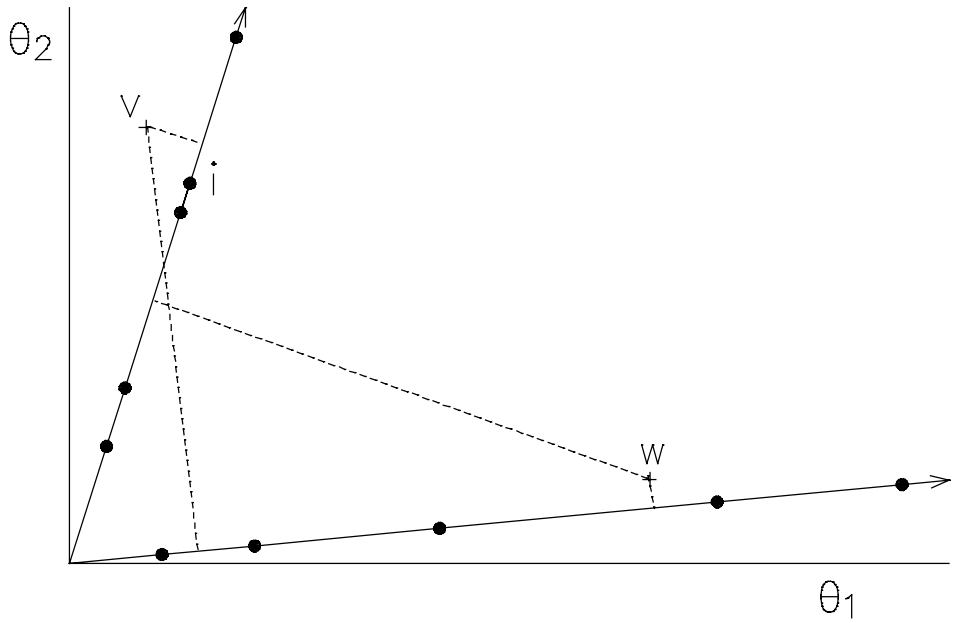
In de psychologie wordt soms gebruikt gemaakt van de Rorschachtest. Daarbij moet de persoon bij tien plaatjes waarop een ongestructureerde inktvlek staat aangeven wat hij of zij in die inktvlek ziet. De antwoorden worden op grond van een theorie uit de persoonlijkheidsleer gecategoriseerd in een aantal categorieën, waarbij ervan wordt uitgegaan dat elke categorie wijst op een bepaalde persoonseigenschap. De kans dat een persoon bij een plaatje een antwoord geeft in een bepaalde categorie zal dus afhangen van de mate waarin deze persoon over de overeenkomstige eigenschap beschikt en van de mate waarin het plaatje een bepaalde categorie van antwoorden uitlokt. Als we de plaatjes beschouwen als items, kunnen we dus stellen dat elk item verschillende latente eigenschappen aanspreekt. Een IRT-model dat het gedrag bij de Rorschachtest adequaat beschrijft, zal dus een multidimensionaal model zijn. In paragraaf 5.5.2 wordt zo’n model besproken.

Een heel andere betekenis van het begrip multidimensionaliteit kan geïllustreerd worden met het volgende voorbeeld. In veel schoolse situaties worden belangrijke beslissingen genomen aan de hand van een enkel rapportcijfer, dat meestal een gewogen gemiddelde is van verschillende proefwerkcijfers. Deze praktijk weerspiegelt de assumptie dat het algemene cijfer, een unidimensionale grootheid, een adequate beslissingsgrond biedt, hoewel niemand zal beweren dat twee leerlingen met hetzelfde cijfer op alle vakken even goed of even slecht zijn. Een slecht cijfer voor wiskunde kan gecompenseerd worden door een goed cijfer voor taal en omgekeerd. Een soortgelijke gedachte kan men van toepassing achten op itemniveau. Als een item een beroep doet

op twee vaardigheden kan een bepaalde kans op een juist antwoord van bijvoorbeeld 0.5 tot stand komen omdat men in beide vaardigheden middelmatig is, maar ook omdat men in de ene vaardigheid erg laag scoort, maar dit tekort kan compenseren omdat men excelleert in de andere vaardigheid. Modellen die dit soort mechanisme veronderstellen worden soms aangeduid als compensatorische modellen. De structuur van deze modellen komt in paragraaf 5.5.3 aan de orde.

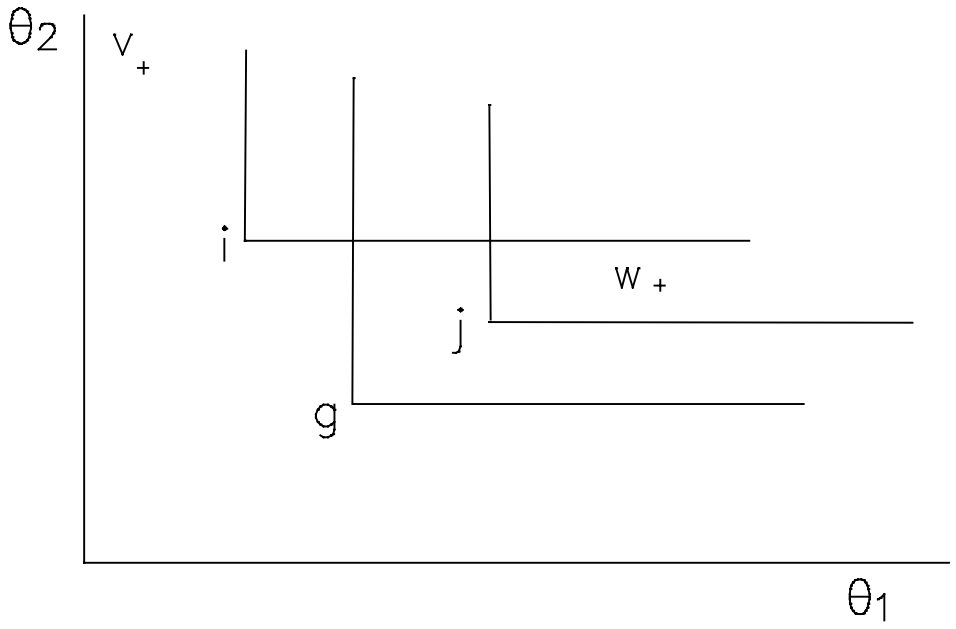
Het voorbeeld van de schoolcijfers is niet helemaal realistisch. De meeste schoolreglementen staan niet toe dat een 1 voor wiskunde gecompenseerd kan worden door een 10 voor taal. Men bouwt dus een mechanisme in de beslissingsregel in, dat bepaalt dat zowel op wiskunde als op taal een bepaald minimum cijfer behaald dient te worden. Dit soort regels kan men ook van toepassing achten op itemniveau. Of een persoon een item juist kan beantwoorden, hangt dan af of een bepaald niveau bereikt is op alle vaardigheden waarop dit item een beroep doet. Modellen die een dergelijk mechanisme veronderstellen worden conjunctieve modellen genoemd. In paragraaf 5.5.4 gaan we op deze modellen in.

De figuren 5.8 en 5.9 zijn een grafisch hulpmiddel om het onderscheid tussen compensatorische en conjunctieve modellen te verduidelijken. Figuur 5.8 is een voorstelling van een compensatorisch model waarbij alle items in de figuur voorgesteld met stippen een beroep doen op de vaardigheden $\theta_1$ en $\theta_2$.




Figuur 5.8
Een compensatorisch model

Vijf items liggen op een lijn die bijna verticaal staat, waarmee wordt aangegeven dat deze vijf items op dezelfde manier een beroep doen op de twee vaardigheden; ze doen echter meer een beroep op $\theta_2$ dan op $\theta_1$, want de hoek die de lijn vormt met de verticale as is kleiner dan de hoek met de horizontale as.




Figuur 5.9
Een conjunctief model

Deze vijf items samen meten dus een unidimensionale vaardigheid, die een bepaald mengsel is van de beide vaardigheden $\theta_1$ en $\theta_2$. De pijl die bij de lijn getekend is geeft de richting van de toenemende vaardigheid aan. Mutatis mutandis geldt dit ook voor de andere vijf items. De tien items samen meten echter niet een unidimensionale vaardigheid, omdat het mengsel van vaardigheden waarop ze een beroep doen niet voor alle items hetzelfde is. De positie van de *letter v* in de figuur geeft aan dat persoon *v* over een hoge mate van vaardigheid $\theta_2$ beschikt, maar over een lage mate van vaardigheid $\theta_1$. We verwachten dus dat die persoon het goed zal doen op items die vooral een beroep doen op $\theta_2$ en minder goed op items die vooral $\theta_1$ aanspreken. Het omgekeerde geldt voor persoon *w*. Om te weten of persoon *v* het goed zal doen bij de beantwoording van item *i*, nemen we de projectie van het punt dat zijn vaardigheid voorstelt op de lijn die de schaal voorstelt waarop het item ligt. We kunnen dit op een analoge manier doen voor de tweede schaal, en ook voor persoon *w*. Deze projecties zijn aangegeven als de eindpunten van de stippellijnen. Met een deterministische interpretatie zouden we kunnen zeggen dat persoon *v* over meer van de gecombineerde vaardigheid beschikt dan item *i* vereist, en dat deze persoon item *i* dus correct zal

beantwoorden. Met deze interpretatie is gemakkelijk uit de figuur af te leiden dat de personen *v* en *w* elk vijf van de tien items juist zullen beantwoorden. Hun scores zijn dus gelijk, hoewel hun begaafdheden drastisch verschillen. Ze hebben beide op een verschillende manier hun tekort op de ene vaardigheid gecompenseerd door een grote mate van de andere vaardigheid.

In figuur 5.9 is een voorstelling van een conjunctief model gegeven. De positie van de items valt samen met het snijpunt van een horizontaal en een verticaal lijnstuk. In een deterministische interpretatie stelt de hoogte van het horizontale lijnstuk de minimale hoeveelheid vaardigheid $\theta_2$ voor die nodig is om het item correct te beantwoorden. Het verticale lijnstuk geeft de minimale hoeveelheid van vaardigheid $\theta_1$ aan. Men kan een item alleen dan juist beantwoorden als men zich rechts boven het punt bevindt dat het item voorstelt. Persoon *v* zal dus geen enkel item juist beantwoorden, en persoon *w* zal een juist antwoord geven op de items *j* en *g*. Hoewel persoon *v* duidelijk over meer vaardigheid $\theta_2$ beschikt dan persoon *w*, helpt dat niet om het tekort aan vaardigheid $\theta_1$ te compenseren.

### *5.5.1 Een OPLM met een multivariate vaardigheidsverdeling*

Indien een unidimensionaal OPLM geen goede passing oplevert, kan men op zoek gaan naar een opdeling van de items in deelverzamelingen die wel goed te beschrijven zijn met een unidimensionaal model. Het zoeken naar zo'n opdeling is geen triviaal probleem en het kan op verschillende manieren gebeuren. Men kan bijvoorbeeld gebruik maken van de toets voor unidimensionaliteit die door Martin-Löf ontwikkeld is (zie hoofdstuk 4), of een factoranalyse uitvoeren op de matrix van interitemcorrelaties (Bol & Verhelst, 1985). Wij gaan niet op dit probleem in. Indien men zo'n opdeling heeft, rijst de vraag hoe de vaardigheden die door de verschillende deeltoetsen worden gemeten met elkaar in verband staan. Een elegante manier om dit probleem aan te pakken, is een multivariate normale verdeling te veronderstellen voor de vaardigheid $\theta = (\theta_1,...,\theta_q,...,\theta_Q)$. Een multivariaat normale verdeling is net als de gewone normale verdeling, eigenlijk een familie van verdelingen, en een lid van deze familie wordt gespecificeerd door de waarden van de parameters vast te leggen. Deze parameters zijn de vector van gemiddelden $\boldsymbol{\mu} = (\mu_1,...,\mu_Q)$ en de covariantiematrix $\Sigma$, waarin niet alleen de variantie van elk van de afzonderlijke $\theta$-variabelen wordt gespecificeerd maar ook hun covarianties. Bij een *Q*-variate normale verdeling zijn er dus $Q + Q(Q+1)/2$ parameters. Indien de oorspronkelijke *k* items zijn opgedeeld in *Q* deelverzamelingen, kan men het nulpunt van de *Q* schalen vrij kiezen, door

bijvoorbeeld alle gemiddelden gelijk te stellen aan 0. In totaal moeten er dus $k + Q(Q+1)/2$ parameters geschat worden.

Als we het antwoordpatroon op de $q$-de deeltoets aanduiden als $\boldsymbol{x}^{(q)}$, en het antwoord-patroon voor alle $k$ items als $\boldsymbol{x} = (\boldsymbol{x}^{(1)},..,\boldsymbol{x}^{(Q)})$, kunnen we de aannemelijkheidsfunctie schrijven als

$$\begin{aligned} L(\beta,\Sigma;\boldsymbol{x}) &= \int\!\!\ldots\!\int P(\boldsymbol{x}\,|\,\theta;\beta)\,g(\theta;\Sigma)\,d\theta \\ &= \int\!\!\ldots\!\int \prod_{q=1}^{Q} P(\boldsymbol{x}^{(q)}\,|\,\theta_q;\beta^{(q)})\,g(\theta;\Sigma)\,d\theta_1\ldots d\theta_Q\,, \end{aligned} \tag{5.49}$$

waarin $\beta^{(q)}$ de vector met itemparameters is voor de items in de $q$-de subtoets. De aannemelijkheidsfunctie gegeven de observaties van verschillende personen is dan gewoon het produkt van uitdrukkingen als het rechterlid van (5.49). Merk op dat (5.49) de multivariate versie is van de aannemelijkheidsfunctie die we in hoofdstuk 4 opgesteld hebben bij de bespreking van de MML-schattingsmethode. In deze context is dit heel natuurlijk, want de toevoeging van een veronderstelling over de verdeling van de vaardigheid in de populatie is een essentieel onderdeel van het model. Glas (1989, 1992) bespreekt de details van de schattingsprocedure en geeft ook aan hoe het model statistisch kan worden getoetst.

Een eenvoudiger versie van dit model werd eerder voorgesteld door Andersen (1985). Bij de toepassing die Andersen bespreekt, levert de opdeling van de items in subtoetsen geen enkel probleem op. Indien dezelfde toets op twee verschillende tijdstippen aan dezelfde personen wordt afgenomen, kan men proberen te achterhalen of en hoe de vaardigheid in de tussentijd is veranderd. Door te veronderstellen dat de verdelingen van $\theta$ op de twee tijdstippen gezamenlijk een bivariaat normale verdeling vormen, krijgt men direct een speciaal geval van het model dat hierboven werd besproken met $Q$=2. Andersen veronderstelde bovendien dat de itemparameters bekend zijn, bijvoorbeeld uit een voorafgaande calibratie. De waarden van de itemparameters op de twee tijdstippen zijn dus exact gelijk. Daarmee liggen de nulpunten van de twee schalen vast, en moeten de gemiddelden $\mu_1$ en $\mu_2$ geschat worden, evenals de twee varianties en de covariantie. Het verschil $\mu_2 - \mu_1$ geeft de gemiddelde toename in vaardigheid, maar het model laat toe dat de twee varianties verschillend kunnen zijn, en dat de correlatie tussen $\theta_1$ en $\theta_2$ ongelijk is aan 1. Men zou kunnen opmerken dat er nooit een correlatie van 1 gevonden wordt tussen twee metingen. Dit is zo, als het gaat over correlaties tussen geobserveerde variabelen die

altijd een zekere mate van onbetrouwbaarheid bevatten waardoor de correlatie niet 1 kan zijn. Hier gaat het echter om de correlatie tussen latente variabelen, die per definitie geen meetfout bevatten. De hoogte van de correlatie geeft een aanduiding van de stabiliteit in de tijd van de latente vaardigheid.

### *5.5.2 Het multidimensionale model van Rasch*

Rasch heeft niet alleen het zeer bekende Raschmodel voor dichotome items ontwikkeld. Hij heeft ook aandacht besteed aan polytome items. In zijn bekommernis om modellen te ontwikkelen waarbij de eigenschappen van items, de itemparameters, bepaald kunnen worden onafhankelijk van wie de items heeft beantwoord, en omgekeerd, waar de eigenschappen van personen gemeten kunnen worden, onafhankelijk van welke items men daar voor gebruikt, kwam Rasch (1961) tot een merkwaardig resultaat: indien de antwoorden op de items in *m* verschillende categorieën kunnen worden ondergebracht, dan hebben we een *m*-dimensionaal model nodig, waarbij de categorieresponscurven gegeven zijn door:

$$P(X_i = j \mid \xi_v) = \frac{\exp(\xi_v^{(j)} - \eta_{ij})}{\sum_{h=1}^{m} \exp(\xi_v^{(h)} - \eta_{ih})}, \quad (j = 1,...,m) \tag{5.50}$$

waarin $\xi_v = (\xi_v^{(1)},...,\xi_v^{(m)})$ en $\xi_v^{(j)}$ geïnterpreteerd kan worden als de mate waarin persoon *v* de neiging heeft om een antwoord in categorie j te geven. Denk hierbij aan de toepassing over de Rorschachtest die we eerder bespraken. De parameter $\eta_{ij}$ kan dan geïnterpreteerd worden als de mate waarin item *i* een antwoord in categorie *j* uitlokt.

Het model dat in (5.50) is gegeven is echter niet geïdentificeerd, omdat er twee soorten transformaties zijn die we op het rechterlid van (5.50) kunnen uitvoeren, zonder dat het linkerlid verandert. Vermenigvuldigen we teller en noemer van (5.50) met $\exp(\eta_{i1} - \xi_v^{(1)})$ en definiëren we

$$\theta_v^{(j)} = \xi_v^{(j)} - \xi_v^{(1)}, \quad (j = 1,...,m), \tag{5.51}$$

$$\beta_{ij} = \eta_{ij} - \eta_{i1}, \quad (j = 1,...,m\,;\, i = 1,...,k), \tag{5.52}$$

dan kan (5.50) herschreven worden als

$$P(X_i = j \mid \theta_v) = \frac{\exp(\theta_v^{(j)} - \beta_{ij})}{1 + \sum_{h=2}^{m} \exp(\theta_v^{(h)} - \beta_{ih})}, \quad (j = 2,...,m) \tag{5.53}$$

en voor het geval $j = 1$ als

$$P(X_i = j \mid \theta_v) = \frac{1}{1 + \sum_{h=2}^{m} \exp(\theta_v^{(h)} - \beta_{ih})}, \quad (j = 2,...,m). \tag{5.54}$$

De 1 in de formules (5.53) en (5.54) verschijnt dus als gevolg van de transformaties (5.51) en (5.52), waaruit direct volgt dat $\theta_v^{(1)} = \beta_{i1} = 0$ voor alle personen $v$ en alle items $i$. Dit betekent dat de neiging om in een bepaalde categorie te antwoorden niet in absolute zin kan worden bepaald. De parameter $\theta_v^{(j)}$ moet dus geïnterpreteerd worden als de sterkte van de neiging om met categorie $j$ te antwoorden vergeleken met de neiging om met categorie 1 te antwoorden. Categorie 1 heet de referentiecategorie. Het blijkt dus dat er maar $m$-1 onafhankelijke dimensies zijn. Stellen we $m$ gelijk aan 2, dan resulteert een unidimensionaal geval, en het is gemakkelijk na te gaan dat in dat geval de formules (5.53) en (5.54) equivalent zijn met de formules voor het unidimensionale Raschmodel dat in hoofdstuk 4 werd behandeld. Merk op dat in dit geval het foute antwoord fungeert als referentiecategorie.

De tweede onbepaaldheid kennen we reeds uit het unidimensionale geval. Indien bij $\theta_v^{(j)}$ en $\beta_{ij}$ een constante $c_j$ opgeteld wordt, verandert hun verschil niet. Dit betekent dat we het nulpunt op elk van de $m$-1 vrije dimensies vrij kunnen kiezen, bijvoorbeeld door $\beta_{1j}$ gelijk te stellen aan 0. Het totale aantal vrije parameters in het model is dus gelijk aan ($k$-1)($m$-1). Hoewel meestal erg makkelijk gedaan wordt over normalisaties, moet men hier toch goed uitkijken, omdat niet alle vergelijkingen van parameters zinvol zijn. De vraag of persoon $v$ meer geneigd is om met categorie $j$ te antwoorden dan persoon $w$, kan men zinvol beantwoorden door het verschil

$$\theta_v^{(j)} - \theta_w^{(j)} = \xi_v^{(j)} - \xi_w^{(j)}$$

te beschouwen. De vraag of persoon *v* meer geneigd is om met categorie *j* te antwoorden dan met categorie *g*, is niet zinvol te beantwoorden, omdat het verschil

$$\theta_v^{(j)} - \theta_v^{(g)}, \quad (j \neq g),$$

volstrekt willekeurig is: de normalisaties van beide dimensies kunnen vrij gekozen worden. Soortgelijke argumenten gelden natuurlijk ook bij het vergelijken van categorieparameters.

Hoewel dit model heel wat eigenschappen heeft die theoretisch zeer aantrekkelijk zijn, waaronder de mogelijkheid om de categorieparameters te schatten met CML, is het bedenken van interessante toepassingsmogelijkheden niet zo eenvoudig. Bovendien is het afleiden van de schattingsvergelijkingen heel wat complexer dan bij het dichotome Raschmodel. De geïnteresseerde lezer kan een gedetailleerde bespreking van de CML-schattingsprocedure vinden in Fischer (1974), waar ook het voorbeeld van de Rorschachtest wordt besproken. Een afleiding van het model vanuit de eis van het bestaan van voldoende steekproefgrootheden voor de persoonsparameters kan men vinden in Andersen (1973c).

### *5.5.3 Compensatorische IRT-modellen*

Uit figuur 5.8 is het vrij gemakkelijk te begrijpen hoe de meeste compensatorische modellen in elkaar zitten. Om de uiteenzetting niet nodeloos ingewikkeld te maken, zullen we de bespreking beperken tot het geval van dichotome items. De gerichte lijn waarop in figuur 5.8 item *i* is afgebeeld kunnen we beschouwen als de reële-getallenas. Het punt dat item *i* voorstelt kan dus geïnterpreteerd worden als een getal, dat we $\beta_i$ zullen noemen. De richting van de lijn is volledig bepaald door de hoeken die de lijn maakt met de twee assen van het assenstelsel, en dus ook door de cosinussen van die hoeken. We duiden die twee cosinussen aan met respectievelijk $a_{i1}$ en $a_{i2}$. Het punt in de tweedimensionale ruimte dat de vaardigheid van persoon *v* aanduidt kunnen we nauwkeurig beschrijven met de twee coördinaten van dat punt, $\theta_{v1}$ en $\theta_{v2}$. De projectie van dit punt op de lijn waarop item *i* ligt is gegeven door

$$a_{i1}\theta_{v1} + a_{i2}\theta_{v2}$$

en dit getal is groter dan $\beta_i$. In de deterministische interpretatie die we eerder gaven, leidde dit positieve verschil tot een juist antwoord. In een kansmodel zullen we zeggen

dat hoe groter dit verschil is, des te groter de kans is op een juist antwoord. Als we gebruik maken van een logistische responsfunctie krijgen we dus automatisch als model:

$$P(X_i = 1 \mid \theta_{v1}, \theta_{v2}) = \frac{\exp(a_{i1}\theta_{v1} + a_{i2}\theta_{v2} - \beta_i)}{1 + \exp(a_{i1}\theta_{v1} + a_{i2}\theta_{v2} - \beta_i)}. \tag{5.55}$$

De generalisatie tot $Q$ dimensies is dan voor de hand liggend:

$$P(X_i = 1 \mid \theta_{v1}, ..., \theta_{vQ}) = \frac{\exp\left(\sum_{q=1}^{Q} a_{iq}\theta_{vq} - \beta_i\right)}{1 + \exp\left(\sum_{q=1}^{Q} a_{iq}\theta_{vq} - \beta_i\right)}. \tag{5.56}$$

Er is echter een eigenschap van het besproken model die nog niet aan de orde is geweest, namelijk dat de som van de kwadraten van de cosinussen $a_{i1}$ en $a_{i2}$ gelijk is aan 1. Deze regel geldt ook indien er meer dan twee dimensies zijn. Dus:

$$\sum_{q=1}^{Q} a_{iq}^2 = 1, \qquad (i = 1, ..., k). \tag{5.57}$$

Uit figuur 5.8 is duidelijk dat, indien we dit model toepassen op de items die allemaal op dezelfde lijn liggen als item *i*, het unidimensionale Raschmodel moet gelden. Dus kan het model dat gedefinieerd is door (5.56) samen met de restrictie (5.57) beschouwd worden als een multidimensionaal compensatorisch Raschmodel. Dit model is in de literatuur echter nog nooit beschreven en bestudeerd. De variant die wel beschreven is, is gegeven door (5.56) waarbij de restrictie (5.57) niet wordt opgelegd (McKinley & Reckase, 1982). De geometrische interpretatie van dit model is iets gecompliceerder dan aangegeven in figuur 4.8, en we gaan er hier niet verder op in; er wordt een interpretatie gegeven in Bol en Verhelst (1985). Als de restrictie (5.57) niet wordt opgelegd, ontstaat een compensatorische generali-satie van het 2PL. Dit is gemakkelijk te zien door in (5.57) $Q$ gelijk te stellen aan 1.

Omdat de gewichten $a_{iq}$ in (5.57) niet bekend zijn, zijn er geen voldoende steekproef-grootheden voor de persoonsparameters, en is CML dus onmogelijk. De schatting van

de parameters gebeurt dan ook meestal met MML, waarbij de veronderstelling gemaakt wordt dat θ $Q$-variaat normaal verdeeld is. Het computerprogramma MAXLOG (McKinley & Reckase, 1983) kan gebruikt worden om de parameters van dit model te schatten.

Lezers die enigszins bekend zijn met factoranalyse, zullen in figuur 5.8 en in de wijze waarop het model is opgebouwd zeker overeenkomsten gezien hebben met de factoranalyse. Als in plaats van de logistische functie, de (cumulatieve) normale verdelingsfunctie als responsfunctie wordt gebruikt en tevens de multivariaat normale verdeling van de vaardigheden, kan aangetoond worden dat het model een uitbreiding is van een factoranalytisch model dat vaak gehanteerd wordt, namelijk het model waarbij de factoren multivariaat normaal verdeeld zijn. Het is een uitbreiding omdat in de factoranalyse alleen de parameters $a_{iq}$ geschat worden, die daar de naam factorlading krijgen, en niet de β-parameters. Bovendien is er een interessant contrast in de manier van parameterschattingen: binnen de traditie van de factoranalyse gebruikt men de correlatiematrix om de parameters te schatten. Indien de variabelen dichotoom zijn, kan deze methode echter tot problemen leiden (zie hoofdstuk 15 van Lord & Novick, 1968). Men kan echter ook de parameters van het model schatten door de aannemelijkheidsfunctie van de geobserveerde antwoordpatronen te maximaliseren, waarbij men meer informatie gebruikt dan aanwezig is in de interitemcorrelatiematrix. De variant van (5.56), waar de normale verdelingsfunctie is gebruikt in plaats van de logistische functie wordt dan ook, met een impliciete referentie naar de schattingsmethode, aangeduid als 'full information factor analysis' (Bock, Gibbons & Muraki, 1988). Het programma TESTFACT (Wilson, Wood & Gibbons, 1991) kan gebruikt worden om de parameters te schatten. Een algemeen overzicht van compensatorische IRT-modellen kan men vinden in Knol (1986).

Tot slot van deze paragraaf komen we nog even terug op een opmerking die in hoofdstuk 4 werd gemaakt, waarin werd betoogd dat het goed mogelijk is dat een unidimensionaal Raschmodel meerdere vaardigheden aanspreekt. Stel dat in figuur 5.8 $\theta_1$ verbale vaardigheid voorstelt, en $\theta_2$ numerieke vaardigheid. Uit de figuur is duidelijk dat alle items beide vaardigheden aanspreken. Als we in een model al deze items betrekken, hebben we inderdaad twee dimensies nodig. Beperken we het model echter tot de items die op dezelfde lijn liggen als item *i*, dan zijn die twee vaardigheden nog wel vereist om deze items te beantwoorden, maar een analyse van de antwoorden zal aanduiden dat we genoeg hebben aan 1 dimensie. Met andere woorden, het 'mengsel' van beide vaardigheden is voor alle items hetzelfde, en we zijn niet meer in staat beide vaardigheden van elkaar te onderscheiden.

### *5.5.4 Conjunctieve IRT-modellen*

Het idee van het stellen van minimumeisen voor verschillende aspecten van een taak is reeds oud (Johnson, 1935), maar in de toegepaste psychometrie zijn de middelen schaars om dit algemene idee op een rationele manier toe te passen. Coombs (1964) heeft er uitvoerig aandacht aan besteed, doch het is pas recent dat er formele modellen zijn ontwikkeld die in de praktijk goed bruikbaar zijn. We bespreken hier kort een model dat door Maris (1992) is ontwikkeld. De deterministische interpretatie van Maris' model is als volgt. Indien aan twee minimumeisen moet worden voldaan, kunnen we ons voorstellen dat er impliciet twee vragen worden gesteld, en het antwoord op het item als geheel is alleen juist indien het antwoord op beide impliciete vragen juist is. Deze impliciete vragen worden natuurlijk niet echt gesteld, en de antwoorden erop zijn dan ook niet observeerbaar. Daarom worden ze latente antwoorden genoemd. Als er $Q$ dimensies zijn, zijn er dus $Q$ latente antwoorden die we zullen aanduiden als $Y_{i1},...,Y_{iQ}$, en die alle de waarden 1 of 0 kunnen aannemen. Het geobserveerde antwoord $X_i$ is alleen gelijk aan 1 indien alle latente antwoorden juist zijn. Het deterministische model kan dus geschreven worden als

$$X_i = \prod_{q=1}^{Q} Y_{iq}. \tag{5.58}$$

Een analyse in het deterministische model komt er dus op neer de items op de $Q$ dimensies zo te ordenen dat alle geobserveerde antwoordpatronen overeenkomen met een gebied in de multidimensionale ruimte dat, onder een conjunctieve interpretatie, met die antwoordpatronen overeenkomt. Zo is er in figuur 5.9 geen plaats voor een antwoordpatroon waarbij alleen item $j$ juist werd beantwoord. Een deterministische oplossing vinden is meestal niet zo eenvoudig, en de reden is, dat het lastig is om te bepalen wat de waarde van $Q$ moet zijn om alle geobserveerde antwoorpatronen hun plaats in de multidimensionale ruimte te geven (Koppen, 1987).

Bij een kansmodel loopt dit iets soepeler omdat in theorie elk antwoordpatroon onder elk model kan voorkomen. Maris construeerde zijn model door aan te nemen dat de latente antwoorden van eenzelfde persoon stochastisch onafhankelijk zijn van elkaar, waardoor we onmiddellijk de probabilistische versie van (5.58) kunnen opschrijven:

$$P(X_i = 1 \mid \theta_1,...,\theta_Q.) = \prod_{q=1}^{Q} P(Y_{iQ} = 1 \mid \theta_q). \tag{5.59}$$

Het model wordt dan gecompleteerd door voor elk latent antwoord het Raschmodel aan te nemen, zodat het model geschreven kan worden als

$$P(X_i = 1 \mid \theta_1,...,\theta_Q) = \prod_{q=1}^{Q} \frac{\exp(\theta_q - \beta_{iq})}{1 + \exp(\theta_q - \beta_{iq})}. \qquad (5.60)$$

Het model is dus een multidimensionaal conjunctief Raschmodel, en we zien dat het unidimensionale Raschmodel resulteert indien $Q = 1$.

De problemen met de parameterschatting en de statistische toetsing van het model zijn zeker niet allemaal opgelost. Zo past Maris alleen de JML-schattingsmethode toe die waarschijnlijk geen consistente schattingen oplevert. Hij beschrijft wel de MML-methode, maar de toepassing ervan brengt vele numerieke problemen met zich mee. Een variant van Maris' model, waarbij wel de MML-methode is gebruikt, kan men vinden in Van Leeuwe (1990).

## 5.6 Nabeschouwing

De grote weelde aan IRT-modellen die in dit hoofdstuk aan bod is gekomen, zal bij de lezer misschien de indruk wekken van wildgroei, zeker als men beseft dat er maar een kleine selectie van de bestaande modellen de revue is gepasseerd. Zie bijvoorbeeld de grote witte oppervlakte rechts en beneden in figuur 5.3. Het grote bos dat men door de vele bomen uit het oog dreigt te verliezen, is bovendien overwoekerd door veel stekelig struikgewas, zoals problemen van statistische, numerieke en algoritmische aard. Het feit dat er een groot aanbod is aan computerprogramma's biedt natuurlijk comfort, doch het zou een misvatting zijn te denken dat de psychometrie bestaat uit een aantal ingewikkelde rekensommen die nu dank zij het beschikbaar zijn van snelle rekenapparatuur gemakkelijk kunnen worden uitgevoerd. De strategie 'ik probeer ze allemaal en ik zie wel welk model het beste past' is een heilloze weg die de verwarring alleen maar groter kan maken. Het toepassen van een psychometrisch model is het toetsen van een hypothese aan de werkelijkheid en deze hypothese dient inhoudelijk zinvol te zijn. Ze probeert de verbanden tussen verschillende gedragingen te formuleren en zo zuinig en accuraat mogelijk te beschrijven. Zie bijvoorbeeld Roskam (1982). De keuze tussen, bijvoorbeeld, een compensatorisch en een conjunctief model moet men niet aan een computerprogramma overlaten, maar baseren op een analyse van het

gedragsdomein dat men wenst te analyseren. De wetenschap dat er goed uitgewerkte psychometrische formaliseringen en bijbehorende computerprogramma's bestaan, wordt dan een bron van welbevinden in plaats van verwarring.