7

Toepassingen van itemresponstheorie

In dit hoofdstuk komen een drietal toepassingen van itemresponstheorie (IRT) aan de orde. Ze zijn enerzijds bedoeld als illustratie van de theoretische uiteenzettingen in de vorige drie hoofdstukken, anderzijds dienen ze om enkele theoretische problemen die niet besproken werden, toe te lichten en een mogelijke oplossing voor te stellen.

De eerste toepassing gaat over een grootschalig Cito-project, de periodieke peiling van het onderwijsniveau (PPON). Het doel van deze peiling is het uitvoeren van metingen en daarover verslag doen. Een van de problemen waarmee het project werd geconfronteerd was het ontbreken van meetinstrumenten. De constructie van de meetinstrumenten en de eigenlijke peiling dienden in één fase te gebeuren. In paragraaf 7.1 worden de psychometrische aspecten van deze dubbele opdracht besproken.

De tweede toepassing behoort tot een domein dat in de psychologie bekend staat als leesbaarheidsonderzoek, een traditie die haar oorsprong vindt in het onderzoek van Vogel en Washburne (1928). De praktische vraagstelling bij dit soort onderzoek betreft de relatie tussen de leesvaardigheid van een jonge lezer en de moeilijkheid of leesbaarheid van een tekst. Met andere woorden, de vraag is of er een maat ontwikkeld kan worden die aangeeft of een bepaalde persoon met goed gevolg een gegeven tekst kan lezen. Hoewel iedereen wel bekend zal zijn met leeftijdscodes op boeken in jeugdbibliotheken, is een dergelijke aanduiding veel te ruw: de spreiding van de leesvaardigheid bij kinderen van dezelfde leeftijd is dermate groot dat deze leeftijdsaanduidingen te enen male onvoldoende zijn. In paragraaf 7.2 worden enkele aspecten van het leesbaarheidsonderzoek van Staphorsius (1992b) besproken.

De derde toepassing heeft betrekking op een beroemde test uit de psychologie, de 'verborgen-figurentest' van Witkin (1950). Met behulp van IRT is door Pennings (1991) een gemodificeerde versie van deze test gemaakt, zodat hij beter geschikt wordt voor diagnostische doeleinden dan de oorspronkelijke test, waarbij alleen aantal juiste antwoorden en gemiddelde antwoordtijd worden geregistreerd. Het is meteen een illustratie van een creatief gebruik van een IRT-model voor polytome items. Deze toepassing wordt in paragraaf 7.3 besproken.

7.1 De PPON-rekenpeiling

In 1987 begon in opdracht van het Ministerie van Onderwijs het project 'Periodieke Peiling van het Onderwijsniveau' (PPON) in het basisonderwijs. Het eerste vakgebied dat werd gepeild was rekenen aan het einde en in het midden van het basisonderwijs, dat wil zeggen bij leerlingen van ongeveer twaalf respectievelijk negen jaar. Het algemene doel van peilingsonderzoek in Nederland kan omschreven worden als: systematisch bijdragen aan het verkrijgen van een beeld van het leeraanbod en de effecten van onderwijs. PPON moet een empirische basis verschaffen voor de algemene maatschappelijke discussie over de inhoud en het niveau van het onderwijs. Concreet betekent dit bijvoorbeeld dat verschillen in leer-prestaties tussen belangrijke subpopulaties in kaart gebracht dienen te worden. De reken-peiling van 1987 is een eerste peiling in een reeks van periodiek herhaalde peilingen, en de resultaten moeten dienen als algemeen referentiepunt om ontwikkelingen in de tijd te kunnen evalueren. Dit aspect van de opdracht, samen met de verplichting om na elke peiling een gedeelte van de items te publiceren, vormt de eerste grote complicatie van de opdracht. De toetsen die gebruikt worden in opeenvolgende peilingen kunnen niet identiek zijn. Dit schept het probleem dat er maatregelen getroffen moeten worden, zodat verschillen in de tijd op gemiddelde prestatie niet ten onrechte kunnen worden toegeschreven aan verschillen in moeilijkheidsgraad.

Een tweede complicerende factor betrof de steekproeftrekking. Omdat het tot de opdracht behoorde betrouwbare en vrij nauwkeurige uitspraken te doen over relatief kleine subpopulaties, bijvoorbeeld etnische minderheden, kon niet worden volstaan met een eenvoudige aselecte steekproef uit de leerlingpopulatie. In dat geval zouden deze minder-heden in onvoldoende aantal in de steekproef vertegenwoordigd zijn. Daarom werd besloten een gestratificeerde steekproef te trekken op zo'n wijze dat scholen met veel leerlingen uit etnische minderheden proportioneel oververtegenwoordigd waren. Bovendien is het om praktische redenen onuitvoerbaar om binnen elk stratum een aselecte steekproef te trekken. Daarom werd gebruikt gemaakt van getrapte steekproeftrekking. Eerst werd uit de populatie van basisscholen een aselecte steekproef getrokken, en dan werd er binnen elke school uit de relevante leeftijdsgroep weer een aselecte steekproef getrokken.

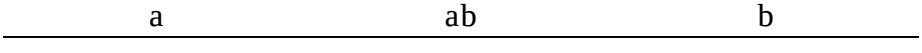
De derde complicatie had te maken met het feit dat de meetinstrumenten nog ontwikkeld moesten worden. Normaliter zou men in een dergelijk grootscheeps onderzoek een constructiefase verwachten waarin de meetinstrumenten ontwikkeld worden, en waarbij een afzonderlijke calibratiesteekproef getrokken wordt om de eigenschappen van het meet-instrument vast te stellen. Door de tijdsdruk bleek dit

echter niet mogelijk te zijn, zodat dezelfde steekproef moest fungeren als calibratiesteekproef en peilingssteekproef, met het theoretische risico dat bepaalde instrumenten van zo'n slechte kwaliteit konden blijken te zijn, dat er van peiling geen sprake meer zou zijn. Bovendien speelden nog andere aspecten van tijdsdruk mee: men kan leerlingen niet een willekeurig lange tijd items laten beantwoorden, en men kan de steekproef niet willekeurig groot maken, wil men de dataverzameling in een realistische periode afronden.

Om een gedetailleerde verslaglegging toe te laten, werd besloten het hele vakgebied rekenen op te delen in inhoudelijk zeer homogene deelgebieden, en voor elk deelgebied een afzonderlijke schaal te construeren. Zo werd bijvoorbeeld het onderwerp 'breuken' opgedeeld in de schalen 'optellen en aftrekken' en 'vermenigvuldigen en delen'. In totaal werden 27 deelgebieden onderscheiden voor de 12-jarigen en 13 deelgebieden voor de 9-jarigen. Voor een gedetailleerde onderwijskundige verantwoording van deze opdeling, zie Wijnstra (1988). Deze opdeling is natuurlijk een gelukkige omstandigheid om het werken met unidimensionale IRT-modellen aanvaardbaar te maken.

De verdere uiteenzetting heeft betrekking op de constructie van één schaal voor één deelgebied. Aan het einde van deze paragraaf komen we nog even terug op de vraagstukken die te maken hebben met het tegelijkertijd hanteren van meer schalen.

In hoofdstuk 4 is het begrip informatiefunctie uiteengezet, waarbij beklemtoond werd dat itemantwoorden niet altijd evenveel informatie geven over de onderliggende vaardigheid. Voor een praktische toepassing als PPON betekent dit dat het nutteloos is hele moeilijke items door hele zwakke leerlingen en zeer gemakkelijke items door hele vaardige leerlingen te laten beantwoorden, omdat die antwoorden nauwelijks informatie opleveren voor het schatten van de itemparameters of de individuele vaardigheid. Om het verzamelen van nutteloze gegevens zoveel mogelijk te vermijden, werd tot de volgende proefopzet besloten. Op grond van het oordeel van de leerkracht, en enkele objectieve criteria zoals het niveau van het geplande vervolgonderwijs, werden alle leerlingen die aan de peiling deelnamen toegewezen aan één van twee niveaugroepen, verder aan te duiden als A en B, waarbij B als vaardiger werd beoordeeld dan A. Merk op dat de groepsindeling slechts één keer plaats vond, en gebruikt werd voor elk van de schalen die de leerlingen beantwoordden. Door de itemconstructeurs werden de items die voor de schaal werden ontwikkeld, ingedeeld in drie niveaus: a voor de gemakkelijke, b voor de moeilijke en ab voor de middelmatig moeilijke items. Het afnamedesign dat werd gebruikt is weergegeven in figuur 7.1. Het betreft dus een onvolledig, groepsgericht design (zie hoofdstuk 6).



Figuur 7.1
Design in het PPON-onderzoek

De designvariabele, het al dan niet aanbieden van een item, is afhankelijk van de schatting van het niveau door de leerkracht, waarbij het aannemelijk is dat deze schatting enige validiteit heeft voor de latente variabele die door de items wordt gemeten, maar anderzijds weer niet volledig samenvalt met de antwoorden op de items die wel zijn aangeboden. Het al dan niet aanbieden van bepaalde items is dus niet volledig bepaald door de geobserveerde itemantwoorden, maar is ook afhankelijk van een variabele die correleert met de niet geobserveerde antwoorden. Dit wil zeggen dat de procedure waardoor het design tot stand gekomen is, niet verwaarloosd mag worden bij ML-schattingen van de modelparameters, op straffe van onzuivere en inconsistente schattingen. Zie hoofdstuk 6 voor een theoretische uiteenzetting hierover. Deze vaststelling heeft een paar heel belangrijke implicaties.

Ze betekent in de eerste plaats dat we een model moeten maken waarin niet alleen de kansen beschreven worden op een goed antwoord, gegeven dat het item aangeboden wordt, zoals bijvoorbeeld het Raschmodel, maar dat we tevens de kansen moeten beschrijven dat een bepaalde leerling, met een bepaalde vaardigheid θ, in de A- of B-groep terecht komt. Stel dat we aannemen dat in de totale populatie θ normaal verdeeld is, dan is het niet realistisch aan te nemen dat alle leerlingen met een θ-waarde boven een bepaalde grenswaarde θ_0 aan de B-groep worden toegewezen, en alle andere leerlingen aan de A-groep. Dit zou immers impliceren dat de toewijzingsprocedure foutloos is, dit wil zeggen dat het leerkrachtoordeel perfect betrouwbaar is en perfect valide met betrekking tot θ. Dit betekent dat in het model de grenswaarde θ_0, de betrouwbaarheid en de validiteit van de leerkrachtoordelen moeten worden opgenomen. Bovendien is dit nog maar een grove benadering van de werkelijkheid, want niet alle leerkrachten beoordelen even betrouwbaar en valide. Dus de verschillen tussen leerkrachten zouden eigenlijk ook gemodelleerd moeten worden.

De tweede implicatie heeft te maken met de wijze van steekproeftrekken. Zelfs al is de veronderstelling waar dat de vaardigheid in de populatie normaal verdeeld is, dan kunnen we dit niet zonder meer gaan invoeren als een modelveronderstelling, omdat de steekproef niet aselect uit de populatie is getrokken. Er moet minstens een model gehanteerd worden voor elk stratum dat voor de steekproeftrekking is gedefinieerd.

Willen we standaard ML-schattingen gaan toepassen, dan zijn we dus verplicht een zeer complex model te gaan ontwikkelen. Nu zou men kunnen redeneren dat al die

argumenten betrekking hebben op de marginale verdeling van θ, en aangezien itemparameterschattingen met MML robuust zijn tegen schendingen van de normaliteitsassumptie (zie het voorbeeld in paragraaf 4.3.6), het niet veel zal uitmaken als we MML-schattingen maken met de modelaanname van één enkele normale verdeling. Jammer genoeg is in dit geval het model niet robuust genoeg, en treden er belangrijke vervormingen op in de schattingen van de itemparameters: de moeilijkheid van de moeilijke b-items wordt systematisch onderschat en die van de gemakkelijke a-items wordt systematisch overschat (Eggen, 1990).

Iets algemener geformuleerd komt het hele probleem erop neer dat we voor de constructie van een meetinstrument opgezadeld worden met een aantal netelige bijkomende problemen die in feite niets met de validiteit van het meetinstrument te maken hebben, maar wel met de verdeling in de populatie van de latente variabele die we met het meetinstrument willen gaan meten. Men zou kunnen opperen dat de onderzoekers, door zo'n ingewikkelde proefopzet te kiezen, dit probleem grotendeels aan zichzelf te wijten hebben. Echter, met een eenvoudige proefopzet is het probleem niet opgelost. Stel dat er een enkelvoudige aselecte steekproef uit de populatie was getrokken, en dat alleen de eenvoudige vraag moest worden beantwoord of jongens gemiddeld meer, minder of evenveel presteren als meisjes, waarbij echter ook in de toekomst moet kunnen worden nagegaan of een eventueel verschil met de tijd toeneemt of afneemt. Door gebruik te maken van een MML-schattingsprocedure om de itemparameters te schatten zijn we verplicht vooraf, per hypothese, een standpunt in te nemen over de structuur van de latente variabele in de populatie. Indien we geloven dat er geen verschil is, kunnen we volstaan met de assumptie van één normale verdeling. Denken we echter dat er verschil zal zijn dan dienen we een verschillende verdeling aan te nemen voor jongens en voor meisjes. Door het invoeren van een hypothese over de verdeling van de latente vaardigheid worden meetprobleem (de eigenschappen van het meetinstrument) en het structurele probleem (de verdeling van de vaardigheid in de populatie) in één samengesteld model met elkaar vermengd. En de grote problemen duiken op indien het model, als geheel, verworpen dient te worden, omdat het statistische toetsingsarsenaal waarover we beschikken niet garandeert dat er in alle gevallen een scherp onderscheid gemaakt wordt tussen schendingen in de meetcomponent en de structurele component van het model.

Het is natuurlijk een veel comfortabeler positie indien het meetmodel gevalideerd kan worden zonder dat aannamen over het structurele model hoeven te worden gemaakt. Dit is mogelijk indien de parameters die betrekking hebben op het meetmodel met de CML-schattingsmethode kunnen worden geschat. Toen het onderzoek uitgevoerd werd, was echter alleen het Raschmodel beschikbaar als IRT-model waar

CML mogelijk was. Het Raschmodel is echter nogal restrictief door de eis van gelijke discriminatie voor alle items, waardoor bij de constructie van een schaal in veel gevallen tamelijk veel items moeten worden verwijderd. Daarom is OPLM ontwikkeld als een soort compromis. Dit model heeft de flexibiliteit van het tweeparameter-logistische model maar het laat CML-schatting van zijn moeilijkheidsparameters toe. De theorie van OPLM is besproken in hoofdstuk 5. Van de ongeveer 500 items in de 40 schalen van de peiling rekenen moest minder dan vijf procent verwijderd worden op grond van de statistische toetsen die in het OPLM-programma zijn geïmplementeerd.

Wanneer het meetmodel eenmaal geaccepteerd is, kan het meetinstrument gebruikt worden om onderzoek te doen naar structurele vraagstukken. Dit kan op verschillende manieren gebeuren. Om een duidelijk idee te hebben van de werkwijze beperken we ons hier tot het geval van twee achtergrondvariabelen, geslacht (jongen-meisje) en herkomst (Nederlands - buitenlands). Als algemene hypothese nemen we aan dat beide variabelen een deel van de variabiliteit in de leerprestatie verklaren. Stellen we de afhankelijke variabele voor als Y_{vjk}, waarbij de index *v* verwijst naar een individu, de index j naar de subpopulatie van de jongens $(j=1)$ respectievelijk meisjes $(j=2)$ en de index *k* naar de subpopulatie van Nederlanders $(k=1)$ respectievelijk buitenlanders $(k=2)$. Een simpel lineair model is gegeven door

$$Y_{vjk} = \mu + \alpha_j + \beta_k + \varepsilon_{vjk}, \tag{7.1}$$

waarin μ een algemene constante is, α_j het effect van de j-de waarde van de geslachts-variabele, en β_k het effect van de k-de waarde van de herkomstvariabele. De grootheid ε_{vjk} is het zogenaamde residu, en wordt beschouwd als een toevalsvariabele waarvoor een bepaalde verdeling wordt aangenomen. We zullen, in overeenstemming met de gewone veronderstellingen uit de variantie-analyse, aannemen dat alle residuen normaal verdeeld zijn met gemiddelde 0 en variantie σ^2:

$$\varepsilon_{vjk} \sim N\left(0,\sigma^2\right). \tag{7.2}$$

Het model, gegeven door (7.1), is niet geïdentificeerd, omdat voor elke gegeven oplossing een andere gemaakt kan worden door α_j met een willekeurige constante c_1 en β_k met een willekeurige constante c_2 te vermeerderen, en ter zelfder tijd $c_1 + c_2$ van μ af te trekken. Er zijn dus oneindig veel mogelijke oplossingen en willen we zinvol over het model praten dan dienen we een oplossing te kiezen. Dat doen we door wat vaak 'technische restricties' genoemd worden, op te leggen aan de parameters. Wij zullen de restricties zo kiezen dat alle effectparameters die '1' hebben als index gelijk worden gesteld aan 0. Dus

$$\alpha_1 = \beta_1 = 0. \tag{7.3}$$

Merk op dat het gemiddelde van nul voor de residuen ook zo'n technische restrictie is en dat we ook een willekeurige andere waarde voor dit gemiddelde hadden kunnen kiezen. De restricties die we hier gekozen hebben, geven echter een elegante interpretatie aan de parameter μ. Beschouw daartoe de verwachte waarde van Y_{v11}:

$$\begin{aligned} \mathcal{E}(Y_{v11}) &= \mu + \alpha_1 + \beta_1 + \mathcal{E}(\varepsilon_{vjk}) \\ &= \mu + 0 + 0 + 0 = \mu . \end{aligned} \tag{7.4}$$

De parameter μ is dus de verwachte waarde van de afhankelijke variabele voor de subpopulatie waar alle categorieën hun 'eerste' of beter gezegd hun referentiewaarde aannemen. In het voorbeeld is 'jongen' de referentiecategorie voor de variabele 'geslacht' en 'Nederlander' de referentiecategorie voor de variabele 'herkomst'. De parameter μ is dus de gemiddelde θ-waarde van de jongens van Nederlandse herkomst.

Om de modelparameters $(\alpha_2, \beta_2, \sigma^2)$ consistent te schatten is het niet nodig dat de steekproef een aselecte steekproef is uit de totale populatie. De twee achtergrondvariabelen samen delen de totale populatie op in vier subpopulaties, en het is voldoende dat de steekproef uit elke subpopulatie beschouwd kan worden als een aselecte steekproef. De schattings-methode die gebruikt wordt is ML, waarbij de schattingen van de itemparameters uit de calibratiefase als de 'echte' itemparameters, dus als bekende constanten worden behandeld.

Een belangrijke vraag is natuurlijk wat we moeten nemen als de afhankelijke variabele *Y* in (7.1). Als we (7.1) werkelijk als een lineair model voor de vaardigheid θ beschouwen, lijkt het voor de hand te liggen *Y* in (7.1) door θ te vervangen, maar dan hebben we het probleem dat θ latent, dus niet geobserveerd, is. Een mogelijke oplossing is θ te vervangen door een zogenaamde 'proxy', bijvoorbeeld een schatting van θ. De Warm-schatter is een goede kandidaat omdat die schatter voor alle scores bestaat, en bijna zuiver is. Een andere goede kandidaat is de gewogen toetsscore, omdat deze voor niet al te extreme scores een bijna lineaire functie van de Warm-schatter is. Toch kleven aan beide benaderingen een paar nadelen, die men niet moet verwaarlozen.

Het eerste nadeel betreft het verlies aan nauwkeurigheid: de schattingen van θ zijn behept met een schattingsfout. Vullen we in het linkerlid van (7.1) zo'n schatting in, dan moet het residu ε_{vjk} geïnterpreteerd worden als de som van een 'waar' residu, dit wil zeggen, de fout bij het voorspellen van θ uit de predictoren, en de schattingsfout. Daardoor zal de residuele variantie toenemen, maar tevens de standaardfout van de schatters van de regressieparameters μ, α_2, β_2.

Het tweede nadeel heeft te maken met de overblijvende onzuiverheid, en de ongelijke verdeling van die onzuiverheid over de vier subpopulaties. Stel dat in één van

de vier subpopulaties relatief veel perfecte en relatief weinig nulscores voorkomen, dan is de gemiddelde Warm-schatting van de steekproef uit deze subpopulatie een onderschatting van het populatiegemiddelde, en deze onzuiverheid zal ook de schatting van de regressie-parameters beïnvloeden.

Deze twee overwegingen hebben er toe geleid dat in (7.1) toch θ werd ingevuld als afhankelijke variabele. Hoewel θ zelf niet geobserveerd is, hebben we toch informatie over θ via de itemantwoorden. Hierna volgt een korte schets van de schattingsprocedure.

Stellen we het antwoordpatroon van persoon v uit de (j,k)-de subpopulatie voor door $\boldsymbol{x}_{vjk}$ en de bijbehorende score door s_{vjk}, en definiëren we $\lambda = (\mu, \alpha_2, \beta_2, \sigma^2)$, dan kan de aannemelijkheidsfunctie gegeven dit antwoordpatroon, geschreven worden als:

$$\begin{aligned} L(\lambda ; \boldsymbol{x}_{vjk}) &= P(\boldsymbol{x}_{vjk} \mid s_{vjk})\, P(s_{vjk}) \\ &= P(\boldsymbol{x}_{vjk} \mid s_{vjk}) \int_{-\infty}^{+\infty} P(s_{vjk} \mid \theta)\, g_{jk}(\theta;\lambda)\, d\theta\,, \end{aligned} \tag{7.5}$$

waarin $g_{jk}(\theta;\lambda)$ de dichtheidsfunctie is van de verdeling van θ in de (j,k)-de subpopulatie. Het residu ε_{vjk} in het rechterlid van (7.1) is de enige toevalsvariabele, en uit (7.1) en (7.2) volgt dus dat θ_{jk}, dat is de toevalsvariabele θ in de (j,k)-de subpopulatie, normaal verdeeld is met gemiddelde $\mu + \alpha_j + \beta_k$ en variantie σ^2. De eerste factor in het rechterlid van (7.5) is geen functie van de parameters λ, en kan dus behandeld worden als een constante. De aannemelijkheidsfunctie gegeven de itemantwoorden van alle personen samen is het produkt van uitdrukkingen zoals het rechterlid van (7.5), en de ML-schattingen zijn die waarden van de parameters die de aannemelijkheidsfunctie maximaliseren. Een gedetailleerde uiteenzetting van de schattingsprocedure is gegeven in Verhelst en Eggen (1989).

In tabel 7.1 is een voorbeeld gegeven van de effectschattingen van zeven achtergrond-variabelen voor de schaal 'meten en maateenheden' voor de 9-jarigen. De variabele 'stratum' is de stratificatievariabele die gebruikt werd bij het steekproeftrekken, de variabele 'herkomst' geeft aan of de leerling Nederlander (N), dan wel buitenlander (B) was. De variabele 'leertijd' maakt onderscheid tussen kinderen die op het moment van de dataverzameling een kalenderleeftijd hadden van niet meer dan 109 maanden (L), en leerlingen die ouder waren (H). Omdat de data afkomstig zijn van leerlingen die in groep 5, voorheen derde klas, zaten, betreft deze laatste categorie dus leerlingen die één of meer keren gedoubleerd hebben. De variabele 'methode' verwijst naar de gebruikte rekenmethode. Voor de effectschattingen is gebruik gemaakt van de tweedeling Modern-Traditioneel. Categorie '1' van de variabele 'aanbod' verwijst naar leerlingen die, op het moment van de dataverzameling reeds onderwijs hadden

gekregen in de basisprincipes waarop de items een beroep doen. Naast deze variabelen is ook de variabele 'design' opgenomen. Categorie A verwijst naar de kinderen die de 'a' en 'ab' items hebben beantwoord, en categorie B naar de kinderen die de items 'ab' en 'b' voorgelegd kregen. Bij het schatten van de parameters worden de effecten uitgedrukt in de schaal die door de itemparameters is gedefinieerd. In tabel 7.1 is echter een lineaire transformatie toegepast op de schaal, waardoor het geschatte gemiddelde van de totale populatie gelijk is aan 250 en de standaarddeviatie 50. Voor elke variabele is de eerst gerapporteerde categorie gekozen als referentiecategorie. De verhouding z tussen parameter-waarden en standaardfout is bij benadering standaardnormaal verdeeld en kan gebruikt worden als toetsingsgrootheid om voor een parameter α_j de nulhypothese $\alpha_j = 0$ te toetsen. Het is interessant op te merken dat men aan de hand van deze tabel ook enig inzicht kan krijgen in de validiteit van het leerkrachtoordeel: de leerlingen die de moeilijkste items hebben gekregen liggen gemiddeld ongeveer tweederde standaardafwijking boven de kinderen die de gemakkelijke items voorgelegd kregen. Een gedetailleerder onderzoek naar de informatiewinst bij groepsgerichte designs kan men vinden in Verhelst (1989).

Tabel 7.1

Effectschattingen van zeven achtergrondvariabelen
op de schaal 'meten en maateenheden'

Variabele	Cat.	*n*	Eff.	*SE*(eff)	*z*=eff/*SE*
Stratum	1	333	0	---	---
	2	350	-11.49	4.02	-2.86
	3	403	-19.16	4.22	-4.55
Gewicht	N	927	0	---	---
	B	159	-36.72	4.96	-7.40
Geslacht	M	557	0	---	---
	V	529	-7.16	3.19	-2.24
Leertijd	L	902	0	---	---
	H	184	-17.51	4.27	-4.10
Methode	M	654	0	---	---
	T	432	-15.70	3.27	-4.80
Aanbod	1	834	0	---	---
	0	252	-7.59	3.78	-2.01
Design	A	514	0	---	---
	B	572	36.60	3.22	11.36

De effecten in de kolom 'Eff' geven het contrast aan met de referentiecategorie. Het effect van de categorie V van de variabele 'geslacht' bedraagt -7.16 eenheden, dit is ongeveer een zevende deel van de standaardafwijking in de populatie. De geassocieerde *z*-waarde van -2.24 is significant op het 5%-niveau, waarmee wordt aangegeven dat het geslacht, naast de andere variabelen die in de analyse zijn opgenomen, een niet te verwaarlozen effect op de prestatie heeft. Bij de interpretatie van de gerapporteerde contrasten dient men, net als bij de gewone regressie-analyse, zeer voorzichtig te zijn. Uit de tabel volgt niet dat meisjes gemiddeld 7.16 punten lager scoren dan jongens. Het is zelfs mogelijk dat meisjes gemiddeld hoger scoren, zoals uit het volgende fictieve voorbeeld blijkt. Veronderstel dat er slechts twee achtergrond-variabelen van belang zijn, 'geslacht' en 'leertijd', en dat de populatiewaarden van de effecten gelijk zijn aan de geschatte waarden uit tabel 7.1, namelijk -7.16 voor de categorie V van de variabele 'geslacht' en -17.51 voor de categorie H van de variabele 'leertijd'. Veronderstel verder dat de gezamenlijke verdeling van de variabelen 'geslacht' en 'leertijd' overeenkomt met tabel 7.2. Dan is het niet moeilijk na te rekenen dat de gemiddelde θ-waarde van de jongens μ_M gegeven is door

$$\mu_M = [.1(\mu + \alpha_1 + \beta_1) + .4(\mu + \alpha_1 + \beta_2)]/.5$$
$$= [.1(\mu + 0 + 0) + .4(\mu + 0 - 17.51)]/.5 = \mu - 14.008 ,$$

terwijl het populatiegemiddelde van de meisjes,

$$\mu_V = [.4(\mu + \alpha_2 + \beta_1) + .1(\mu + \alpha_2 + \beta_2)]/.5$$
$$= [.4(\mu - 7.16 + 0) + .1(\mu - 7.16 - 17.51)]/.5 = \mu - 10.662 \text{ bedraagt}.$$

Tabel 7.2
Niet-orthogonale verdeling van achtergrondvariabelen, leidend tot Simpsons paradox.

	geslacht	
leertijd	M: $\alpha_1 = 0$	V: $\alpha_2 = -7.16$
L: $\beta_1 = 0$	0.1	0.4
H: $\beta_2 = -17.51$	0.4	0.1

Dus, zowel in de subpopulatie 'leertijd = L' als in de subpopulatie 'leertijd = H' doen de meisjes het minder goed dan de jongens, doch gemiddeld over de hele populatie doen de meisjes het beter. De verklaring van dit paradoxale fenomeen is gelegen in het feit dat beide variabelen, 'geslacht' en 'leertijd' in de populatie niet onafhankelijk zijn,

of zoals men meestal zegt, niet orthogonaal zijn. Dit fenomeen is voor het eerst in de literatuur beschreven door Simpson (1951), en staat bekend als Simpsons paradox. De interpretatie van het geslachtseffect dient dan ook conditioneel te gebeuren: de meisjes scoren gemiddeld 7.16 punten lager dan de jongens indien de andere achtergrondvariabelen constant worden gehouden. Merk op dat de gemiddelde θ-waarde van de jongens of van de meisjes niet uit tabel 7.1 kan worden berekend, omdat de gezamenlijke verdeling van de zeven achtergrond-variabelen niet gegeven is.

Met betrekking tot de standaardfouten dient opgemerkt te worden dat de gerapporteerde getallen een beetje te optimistisch zijn om drie redenen. Ten eerste, de standaardfouten, berekend uit de informatiematrix gelden alleen asymptotisch. In eindige steekproeven zijn de standaardfouten groter. In de tweede plaats is er geen rekening gehouden met het feit dat de itemparameters niet bekend zijn, en dat we ons beholpen hebben met schattingen. Deze schattingen bevatten echter een schattingsfout waarmee geen rekening is gehouden bij het berekenen van de standaardfouten van de regressieparameters. Ten derde is het zo dat de variabelen in tabel 7.1 niet allemaal dezelfde status hebben. De variabelen 'stratum' en 'methode' zijn geen leerlinggebonden variabelen, maar schoolvariabelen. Alle leerlingen in de steekproef die uit dezelfde school komen hebben dezelfde rekenmethode gevolgd. Dit betekent dat, indien 'methode' een effect heeft, de residuen voor leerlingen uit dezelfde school niet onafhankelijk van elkaar zijn. Deze afhankelijkheid is in de analyse veronachtzaamd; er is gedaan alsof alle variabelen leerlinggebonden zijn. Het resultaat is dat de gerapporteerde standaardfouten systematisch te klein zijn. Vergelijk met hoofdstuk 2, de discussie over intraklassecorrelatie. Een correcte analyse zou vereisen dat elke variabele op zijn juiste niveau geanalyseerd wordt. Dergelijke analysemethoden worden aangeduid als multi-niveau- of multi-level-analyses. Er is echter geen programmatuur voorhanden om een multiniveau-analyse uit te voeren waarbij de afhankelijke variabele niet geobserveerd is. Het effect van de fout is, hoewel niet precies bekend, in het geval van de PPON-analyses waarschijnlijk erg klein, omdat de proefopzet zo werd ingericht dat van eenzelfde school niet meer dan vier leerlingen de items van eenzelfde schaal beantwoordden.

Tenslotte zij er nog op gewezen dat de data verzameld zijn in een onvolledige proefopzet, zie figuur 7.1. Voor de schatting van de effectparameters vormt dit geen enkel probleem, omdat in formule (7.5) rekening gehouden wordt met het design, hoewel dat niet expliciet is aangegeven. De factor $P(s_{vjk}|\theta)$ is een functie van de parameters van de items die persoon *v* heeft beantwoord.

7.2 De Cito leesbaarheidsindex voor het basisonderwijs

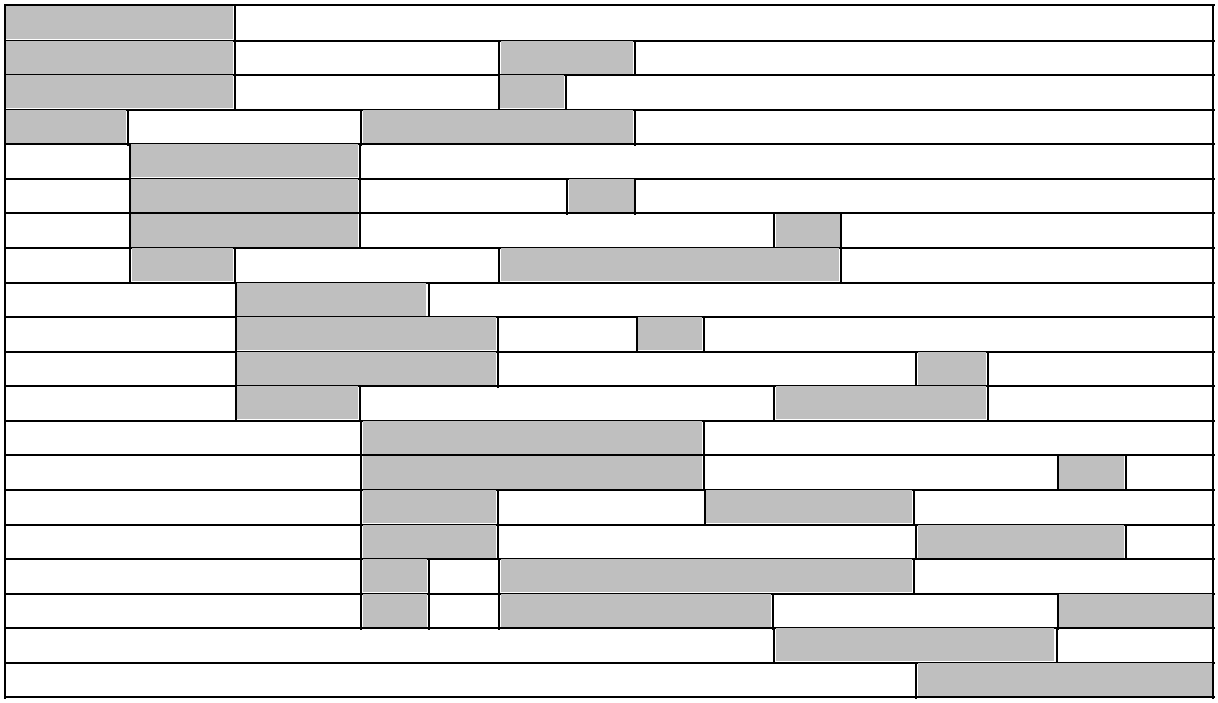
Leesbaarheid

Leesbaarheidsonderzoek heeft in verreweg de meeste gevallen als praktische bedoeling het construeren van een leesbaarheidsindex. Een bruikbare methode hiertoe is de zogenaamde cloze-procedure. Deze procedure bestaat uit het weglaten van woorden uit een tekst volgens een vast patroon. Leerlingen wordt gevraagd de ontbrekende woorden in te vullen. In het te bespreken onderzoek werd elk zevende woord weggelaten, elke tekst heeft zo zeven varianten. Middelen van het aantal correcte antwoorden in een representatieve steekproef over de varianten van de tekst, is nu een maat voor de moeilijkheid van de tekst. Teksten kunnen op deze manier worden geordend naar moeilijkheid. Het is natuurlijk niet praktisch om voor elke nieuwe tekst waarvan men de leesbaarheid wil bepalen deze cloze-procedure toe te passen. Daarom wordt gezocht naar formele tekstkenmerken die in combinatie de gemiddelde score van de tekst goed konden voorspellen. Goede voorspellers zijn onder meer de gemiddelde woordlengte, de gemiddelde zinslengte en het percentage frequente woorden in de tekst. Deze predictoren, die gemakkelijk en betrouwbaar kunnen worden gemeten, worden dan gebruikt als onafhankelijke variabelen in een regressievergelijking. Staphorsius (1992a; maar zie ook Staphorsius & Krom, 1985a en 1985b) vond een multipele correlatie van .85 bij het voorspellen van de gemiddelde cloze-score. De regressie-coëfficiënten die in dit onderzoek zijn gevonden, kunnen dan toegepast worden op willekeurige teksten waarvan de formele kenmerken zijn bepaald. De uitkomst van deze regressieformule, dat wil zeggen de voorspelde gemiddelde cloze-score, wordt de CLIB-waarde van de tekst genoemd. CLIB is de afkorting van Cito leesbaarheidsindex voor het basisonderwijs.

De leesbaarheidsindex van een tekst laat wel toe teksten in moeilijkheid te ordenen, doch hij is niet voldoende om aan te geven of een bepaalde persoon geschikt is voor een gegeven tekst, dat wil zeggen of die persoon de tekst kan lezen en begrijpen. Wat daartoe nodig is, is een maat voor de leesvaardigheid van de persoon en de relatie tussen die lees-vaardigheid en de CLIB-waarde van de tekst. Met andere woorden, we moeten antwoord kunnen geven op de vraag of een leerling met leesvaardigheid x in staat is een tekst met CLIB-waarde y te begrijpen.

Leesvaardigheid

Staphorsius (1992b) heeft een teksttoets ontwikkeld waarbij gebruik werd gemaakt van IRT. De items van de toets bestaan uit tekstfragmenten waaruit een of meer woorden zijn weggelaten. De leerlingen moeten het fragment completeren door uit vijf gegeven antwoordalternatieven het juiste te kiezen. De items zijn zo geconstrueerd dat het juiste antwoord alleen gevonden kan worden indien de tekst die voorafgaat aan en volgt op het ontbrekende stuk, is begrepen. In totaal werden 42 teksten gebruikt die werden opgedeeld in zes fragmenten van ongeveer 180 woorden, zodat er in totaal meer dan 250 items waren. Het spreekt vanzelf dat niet alle items aan eenzelfde persoon ter beantwoording konden worden aangeboden. Het hele onderzoek had betrekking op leerlingen van groep 4 tot en met groep 8 en de variatie in de moeilijkheid van de teksten was voldoende groot om bij het toewijzen van de teksten rekening te kunnen houden met verschillen in leesvaardigheid tussen de leerlingen. Aldus ontstond een onvolledig design dat in principe dezelfde structuur had als het design in figuur 7.1. Het was iets gecompliceerder, omdat de dataverzameling zich over verschillende jaren uitstrekte, zodat een aantal leerlingen gedurende hun hele schoolloopbaan gevolgd kon worden. Een gedeelte van het uiteindelijk gerealiseerde design is afgebeeld in figuur 7.2. De rijen in de figuur komen overeen met groepen leerlingen, geordend volgens geschat leesniveau; de kolommen komen overeen met items geordend volgens geschat moeilijkheidsniveau. In totaal werden meer dan 20.000 antwoordpatronen verzameld, waarbij elk antwoordpatroon de antwoorden bevatte op tussen de 30 en 60 items. Het aantal leerlingen dat aan het onderzoek deelnam was beduidend minder omdat een behoorlijk aantal leerlingen verschillende keren aan de testafname, met gedeeltelijk andere items, deelnam. Elk item werd minimaal 850 keer beantwoord.

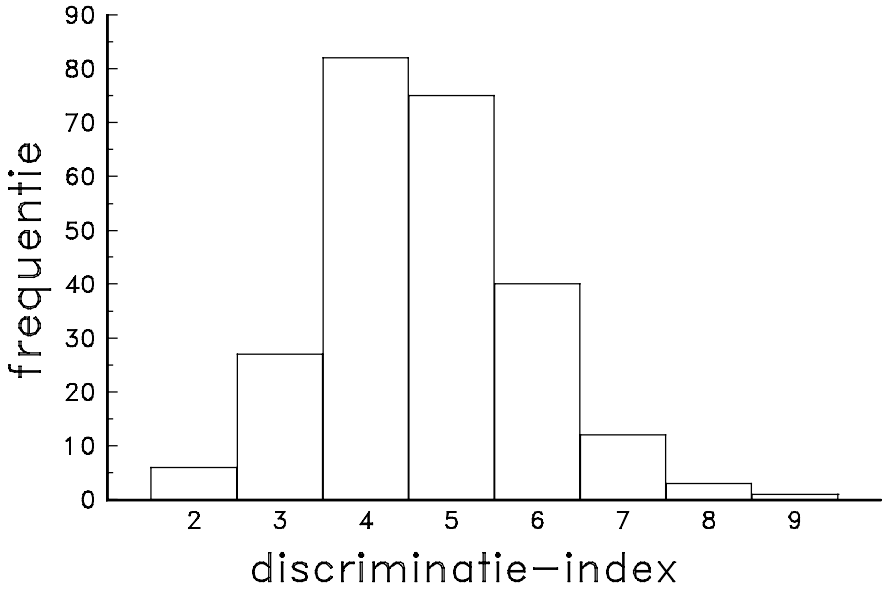


Figuur 7.2
Design van het leesvaardigheidsonderzoek

Net zoals in het PPON-onderzoek vereist een schattingsmethode met MML een vrij gecompliceerd model waarin de designvariabelen, het al dan niet aanbieden van items, gemodelleerd worden. Bovendien treedt hier een extra complicatie op, omdat de steekproeven, overeenkomend met de rijen van figuur 7.2 niet onafhankelijk zijn van elkaar. Verschillende leerlingen namen meer keren aan het onderzoek deel, en deze afhankelijkheid dient gemodelleerd te worden wil men een correcte MML-procedure toepassen. Wordt daarentegen met een CML-procedure gewerkt, dan spelen deze overwegingen geen rol, en ook niet het feit dat leerlingen meermaals aan de test deelnamen. Immers, het is aannemelijk dat na een tussenperiode van een jaar de leesvaardigheid θ veranderd is, en voor het model maakt het niets uit of die twee verschillende θ-waarden afkomstig zijn van één dan wel van twee personen. Voor de verdeling van θ maakt het wel uit: de θ-waarden van twee aselect uit de populatie getrokken personen zijn per definitie onafhankelijk van elkaar, terwijl de θ-waarde van dezelfde persoon op twee verschillende tijdstippen dat niet zijn; dat kunnen we althans niet veronderstellen, anders zou het hele onderzoek zinloos worden.

Het schatten van de itemparameters werd uitgevoerd met het programma OPLM, waarbij de discriminatie-indices een aantal keren werden aangepast. In de uiteindelijke oplossing werden 246 items opgenomen. De verdeling van de discriminatie-indices is afgebeeld in figuur 7.3. Bedenk dat de absolute waarden van deze indices onbelangrijk

zijn, alleen hun onderlinge verhoudingen zeggen iets over het relatieve discriminerende vermogen. Uit de figuur blijkt heel duidelijk dat voor het merendeel van de items de paarsgewijze verhoudingen tamelijk dicht bij 1 liggen, maar toch weer verschillend genoeg zijn om het Raschmodel niet als nulhypothese te kunnen handhaven.

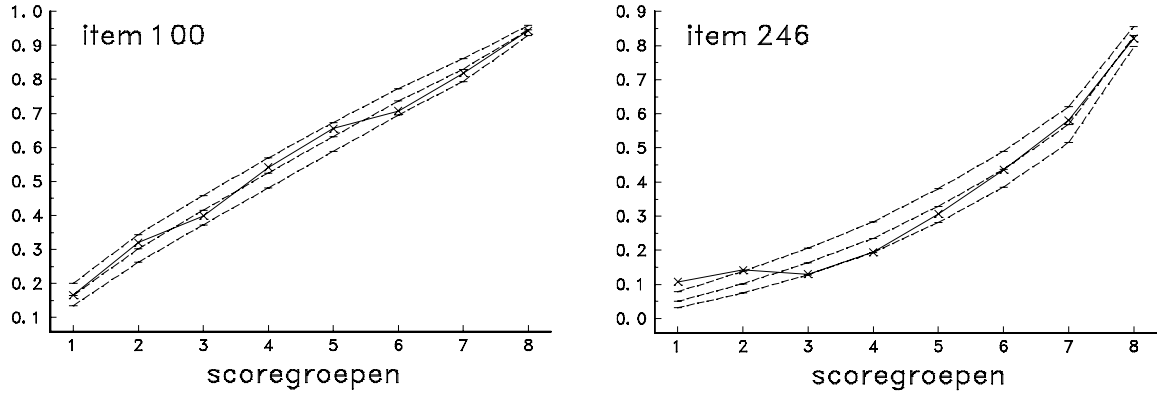


Figuur 7.3
Discriminatie-indices van de 246 items in het leesonderzoek

Om een indruk te geven van de passing van het model, zijn de gegevens waarop de S_i-toetsen gebaseerd zijn, afgebeeld in figuur 7.4 voor twee items. De volle lijnen verbonden door x-symbolen geven de geobserveerde proporties juiste antwoorden weer voor het item, de middelste stippellijn verbindt de voorspelde proporties, en de twee buitenste lijnen geven bij benadering het 95%-betrouwbaarheidsinterval aan. Het item dat links is afgebeeld is een typisch voorbeeld van de meeste items die in de schaal werden opgenomen. Het is bovendien een item dat niet al te moeilijk is: in de hoogst scorende groep is de proportie correcte antwoorden ongeveer 0.9. Het item dat rechts is afgebeeld is het slechtst passende item, en de afbeelding laat meteen ook zien wat de reden van deze slechte passing is. Het is een moeilijk item, en de twee laagst scorende groepen scoren duidelijk hoger dan door het model wordt voorspeld. Dit zou een effect kunnen zijn van het raden bij meerkeuzevragen.

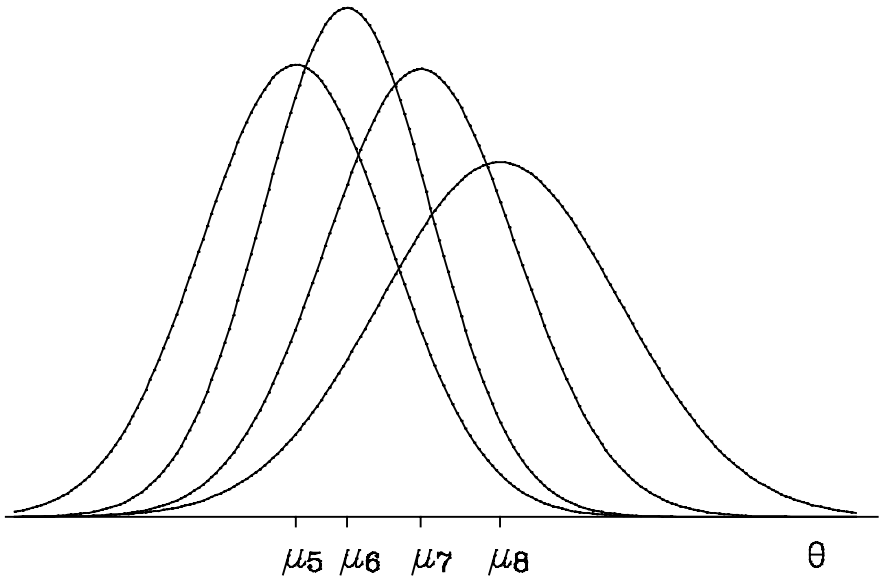
De beoordeling van de algemene modelpassing is een lastig probleem in dit onderzoek: door het zeer groot aantal observaties krijgen de statistische toetsen zeer veel onderscheidend vermogen. Effecten als weergegeven in het rechter gedeelte van figuur 7.4, zelfs als ze veel minder uitgesproken zijn, kunnen bij dergelijke steekproefgrootte gemakkelijk tot significantie aanleiding geven. De procedure van

Hommel die in hoofdstuk 4 is besproken, leidde tot verwerping van het model op het 1%-niveau. Door verwijdering van het slechtst passende item was Hommels toets echter niet significant op het 5%-niveau.



Figuur 7.4
Modelpassing van twee items uit het leesbaarheidsonderzoek

Om een idee te krijgen van de verdeling van de leesvaardigheid in de verschillende jaargroepen werden uit de totale steekproef vier deelsteekproeven gebruikt die representatief konden worden geacht voor de vier onderscheiden populaties, de groepen 5 tot 8. Elke steekproef bevatte ongeveer 1200 leerlingen. In totaal waren er 219 items door de vier deelsteekproeven gemaakt. Op analoge wijze als in paragraaf 7.1 werd beschreven, werden van elke populatie het gemiddelde en de standaardafwijking geschat. Een grafische weergave van de resultaten is gegeven in figuur 7.5.



Figuur 7.5

Verdeling van de leesvaardigheid voor de jaargroepen 5 tot 8

Uit de figuur blijkt zeer duidelijk dat de variabiliteit van de leesvaardigheid groot is in vergelijking met de spreiding tussen de gemiddelden μ_i van de respectievelijke jaargroepen. Dit geeft achteraf gezien een bevestiging van de zinvolheid van het onderzoek: alleen een jaargroep aangeven als indicatie voor de geschiktheid van lectuur negeert de variabiliteit binnen de jaargroepen. De variantie tussen de jaargroepen bedraagt 38% van de totale variantie. Dit betekent dat, indien de jaargroep beschouwd wordt als een maat van lees-vaardigheid, dat wil zeggen een één-item toets, deze een betrouwbaarheid heeft van .38 met betrekking tot de totale populatie van 5- tot 8-jarigen. De uiteindelijk geconstrueerde toetsen (Staphorsius, 1992b) die nu in het onderwijs worden gebruikt, hebben een betrouwbaarheid van boven de .95 met betrekking tot dezelfde populatie, en verklaren dus meer dan 95% van de variabiliteit.

Validiteit

Bij het gebruik van een IRT-model, gaat men uit van bepaalde axioma's, en de statistische toetsen worden gebruikt om de aanvaardbaarheid van deze axioma's te toetsen. Deze toetsen maken dus deel uit van het valideringsonderzoek. Doch daarmee is het valideringsonderzoek natuurlijk niet afgelopen, enerzijds omdat er modelschendingen kunnen zijn die de statistische toetsen niet ontdekken, anderzijds omdat er aspecten zijn aan valideringsonderzoek waarvoor de gebruikelijke statistische modeltoetsen niet geschikt zijn. Er is bijvoorbeeld geen enkele mogelijkheid om uit alleen de leesvaardigheidsdata het besluit te trekken dat de items leesvaardigheid en niet iets anders meten. Voor dit aspect van de validiteit hebben we een extern criterium nodig. We bespreken eerst een bijkomende manier om de geldigheid van het model te controleren, en vervolgens gaan we in op een aspect van de criteriumvaliditeit.

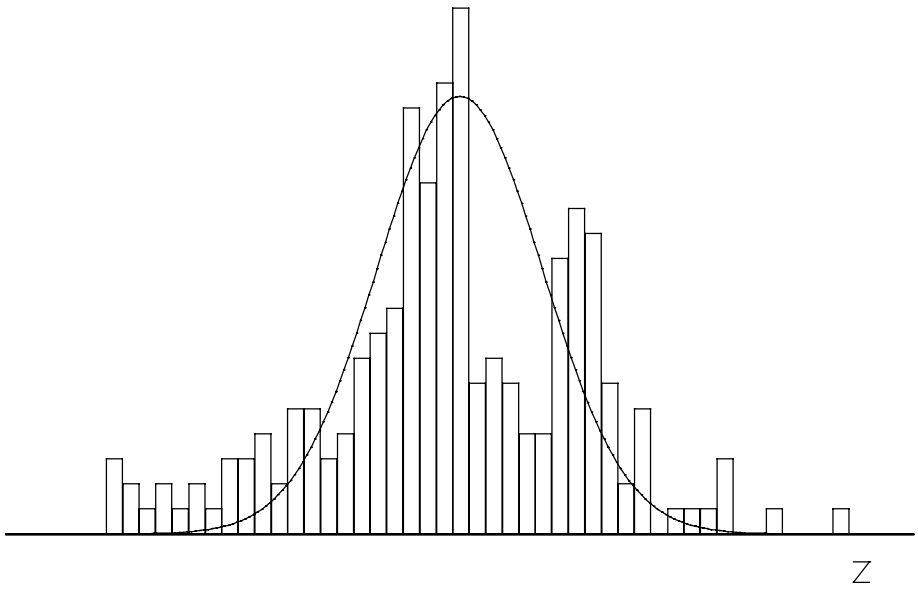
In de klassieke testtheorie wordt de moeilijkheid van een item doorgaans aangegeven met zijn theoretische *p*-waarde, de kans dat het item door een aselect getrokken persoon uit de populatie juist wordt beantwoord. De proportie juiste antwoorden in de steekproef is een schatting van de theoretische *p*-waarde, die we zullen aanduiden als π_i voor item *i*. Indien een IRT-model geldig is, met itemresponsfuncties $f_i(\theta)$, en de verdeling van θ in een bepaalde populatie is gegeven door de dichtheidsfunctie $g(\theta)$, dan geldt dat

$$\pi_i = \int_{-\infty}^{+\infty} f_i(\theta) g(\theta) d(\theta). \qquad (7.6)$$

Zowel $f_i(\theta)$ als $g(\theta)$ is een functie van de modelparameters. Vullen we in die functies nu schattingen van de parameters in, dan is het rechterlid van (7.6) een schatter van π_i, die niet noodzakelijkerwijze precies moet gelijk zijn aan de proportie juiste antwoorden, omdat de data die hier gebruikt worden een deelverzameling zijn van de data waaruit de itemparameters zijn geschat. Maar het verschil tussen beide schatters: $\hat{\pi}_i$, berekend door in het rechterlid van (7.6) de schattingen van de parameters in te vullen, en de geobserveerde proportie p_i, mag niet al te groot zijn, want beide zijn consistente schatters van dezelfde grootheid π_i. Voor alle items die gebruikt werden bij het schatten van de verdelingen in de jaargroepen 5 tot 8 zijn beide grootheden uitgerekend. In figuur 7.6 is het histogram van de gestandaardiseerde afwijkingen

$$z_{(p_i - \hat{\pi}_i)} = \frac{\sqrt{n_i}\,(p_i - \hat{\pi}_i)}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}, \tag{7.7}$$

gegeven, waarbij n_i het aantal personen is dat item i heeft gemaakt. De gestandaardiseerde afwijkingen, gegeven door (7.7) zijn bij benadering normaal verdeeld met gemiddelde 0. De standaardafwijking is echter niet gelijk aan 1, omdat geen rekening is gehouden met het feit dat $\hat{\pi}_i$ niet de werkelijke parameter is, doch een schatting. Omdat de calibratiesteekproef zo groot is, zal het effect van deze fout waarschijnlijk niet al te groot zijn. Het effect van deze verwaarlozing van de schattingsfout maakt dat de gestandaardiseerde afwijkingen gegeven in (7.7) een standaardafwijking hebben die groter is dan 1. Om toch enige indruk te krijgen van de passing van het model is een standaardnormale verdeling bij het histogram getekend.



Figuur 7.6
Gestandaardiseerde afwijkingen tussen geobserveerde en voorspelde proporties

Zelfs al is de standaardafwijking van de theoretische verdeling onderschat, dan blijkt uit de figuur nog heel duidelijk een relatief te groot aantal negatieve z-waarden met grote absolute waarde, terwijl afwijkingen met kleine positieve waarden niet vaak genoeg voorkomen. Een negatieve *z*-waarde betekent dat de voorspelde waarde $\hat{\pi}_i$ groter is dan de geobserveerde proportie p_i. Een verklaring voor dit effect ligt wellicht wederom in raadgedrag als gevolg van het gebruik van meerkeuzevragen. Het item dat in figuur 7.4 rechts is afgebeeld, leverde de kleinste *z*-waarde op ($z = -4.23$). Uit de figuur blijkt het raadgedrag duidelijk bij de twee laagste scoregroepen, doch dit betekent natuurlijk niet dat raadgedrag tot die twee groepen beperkt is gebleven. Men kan geredelijk aannemen dat er ook geraden is, hoewel in mindere mate, in de andere scoregroepen. Bij de schatting van de itemparameters wordt de geobserveerde proportie juist gelijkgesteld aan de voorspelde proportie, dat wil zeggen, het item wordt gemakkelijker geschat dan het werkelijk is, omdat een gedeelte van de juiste antwoorden is toe te schrijven aan raden en niet aan voldoende leesvaardigheid. Dit heeft dan als gevolg dat er een systematische fout in de itemparameterschattingen wordt geïntroduceerd, die op haar beurt doorwerkt in de schatting van de populatieparameters. Of hierin inderdaad een voldoende verklaring ligt voor de afwijkingen is echter niet helemaal duidelijk, en dient onderwerp te zijn van verder onderzoek.

Wij volstaan hier met een algemene beschouwing, die aansluit op wat in hoofdstuk 4 werd gesteld. Het gebruik van het Raschmodel of van een ander model dat CML-schattingen toelaat, heeft het grote voordeel van de zogenaamde steekproefonafhankelijkheid, waarbij het er niet toe doet hoe de steekproef uit de populatie is getrokken. In het onderzoek van Staphorsius is van dit voordeel op grote schaal gebruik gemaakt: de totale steekproef waarop de calibratie is uitgevoerd, getuigt op het eerste gezicht van een soort wildgroei, die elke poging om tot een min of meer realistische beschrijving van de verdeling van θ bij voorbaat tot een hopeloze onderneming maakt. De ingewikkeldheid van het design heeft echter zijn redenen, omdat veel data werden verzameld met andere doeleinden dan alleen het toepassen van een meetmodel. Het verzamelen van herhaalde metingen bij dezelfde personen bijvoorbeeld heeft geleid tot het inpassen van dit onderzoek in het leerlingvolgsysteem dat op het Cito is ontwikkeld. Het grote voordeel van de steekproefonafhankelijkheid kan echter alleen geclaimd worden indien het meetmodel geldig is. Indien meerkeuzevragen gebruikt worden, en er wordt in meer of mindere mate geraden, dan verdwijnt dit voordeel. Zelfs bij redelijk goed uitvallende modeltoetsen, zoals bij de data van Staphorsius, treden er systematische fouten op zodra het model wordt toegepast op populaties die systematisch verschillen van de populatie die bij de

calibratie werd gebruikt, zoals uit figuur 7.6 blijkt. Dit betekent natuurlijk niet dat de onderzoeksgegevens van Staphorsius onbruikbaar zijn. Bij 90% van de items is het absolute verschil tussen geobserveerde en voorspelde *p*-waarde kleiner dan 0.035, en bij 80% is het kleiner dan 0.02. De praktische consequenties zijn tweevoudig: ten eerste kan het toepassen van de geconstrueerde schaal leiden tot een verkeerde schatting van verschillen tussen populaties waar het raadgedrag systematisch gaat verschillen; ten tweede levert het gebruik van meerkeuze-items in modellen die niet voorzien in raadgedrag, dus andere modellen dan bijvoorbeeld het drieparametermodel, bijna automatisch de hierboven beschreven problemen op. Hoewel op het eerste gezicht het gebruik van dit soort ingewikkelder modellen voor de hand schijnt te liggen, is de CML-schattingsmethode hierbij uitgesloten, en is men bij ingewikkelde designs aangewezen op een zeer ingewikkelde modellering van de verdeling van θ, waarbij men zich vaak tevreden zal moeten stellen met benaderingen waarvan het allerminst zeker is of ze een even goede predictie opleveren als in figuur 7.6 is afgebeeld. Een suggestie die vanuit psychometrisch oogpunt voor de hand lijkt te liggen, namelijk afzien van meerkeuze-items, lijkt de oplossing van het probleem te zijn. Voor de praktische haalbaarheid van deze oplossing zal het oordeel van de veldonderzoeker wellicht zwaarder moeten wegen dan een suggestie uit de psychometrie.

Voor het tweede onderdeel van de validiteitsstudie, namelijk de relatie met externe variabelen, beperken we ons tot één gedeelte uit het onderzoek van Staphorsius. Indien de teksttoets dezelfde vaardigheid meet als een cloze-toets, dan bestaat de voor de hand liggende controle erin, de teksten van de teksttoetsen te 'be-clozen' en het verband na te gaan tussen individuele cloze-scores en de geschatte vaardigheid θ die door de teksttoets wordt gemeten. De dataverzameling voor dit doel is begonnen, doch bij het schrijven van dit hoofdstuk waren de resultaten nog niet beschikbaar. Toch kunnen we indirecte evidentie voor dit verband krijgen door de $\hat{\pi}_i$-waarden die met (7.6) te berekenen zijn, te beschouwen als 'proxies' voor de cloze-scores. Van alle 246 items werd de gemiddelde $\hat{\pi}_i$-waarde berekend over de jaar-groepen 5 tot 8. Om de overeenkomst met de cloze-procedure te bevorderen, werden de $\hat{\pi}_i$- waarden van items die tot dezelfde tekst behoren, gemiddeld en beschouwd als 'proxy' voor de cloze-scores. Indien de teksttoets dezelfde vaardigheid meet als de cloze-score, dan moet de voorspelling van de gemiddelde $\hat{\pi}_i$-waarden uit formele tekstkenmerken goed overeenkomen met de CLIB-waarde van die teksten. De multipele correlatie tussen de gemiddelde $\hat{\pi}_i$-waarden en formele tekstkenmerken bedroeg .967. Het feit dat deze correlatie hoger is dan de correlatie tussen deze formele tekstkenmerken en de gemiddelde cloze-scores, is voor een deel te verklaren uit het feit dat de gemiddelde $\hat{\pi}_i$-waarden een grotere spreiding vertonen dan de gemiddelde cloze-scores. Bovendien

waren de teksten waarop de cloze-scores zijn bepaald, een steekproef uit bestaande teksten, waarvan sommige zeer specifieke kennis vereisten en zodoende de cloze-score drukten. Bij het formuleren van de teksttoetsen daar-entegen was veel zorg besteed om de antwoorden zoveel mogelijk onafhankelijk te maken van specifieke kennis of informatie die niet in de tekst gegeven was. De hoge correlaties tussen enerzijds cloze-score en formele tekstkenmerken, en anderzijds tussen gemiddelde $\hat{\pi}_i$-waarden en formele tekstkenmerken, impliceren een hoge correlatie tussen gemiddelde cloze-score en gemiddelde $\hat{\pi}_i$-waarden. De correlatie tussen de voorspelde waarde van de gemiddelde $\hat{\pi}_i$-waarden en de CLIB bedroeg 0.987.

De correlatie tussen individuele cloze-scores en de geschatte θ-waarde zal ongetwijfeld lager uitvallen; maar niettemin zijn deze resultaten duidelijke evidentie dat teksttoetsen en cloze-toetsen dezelfde vaardigheid aanspreken.

Het verband tussen leesvaardigheid en leesbaarheid

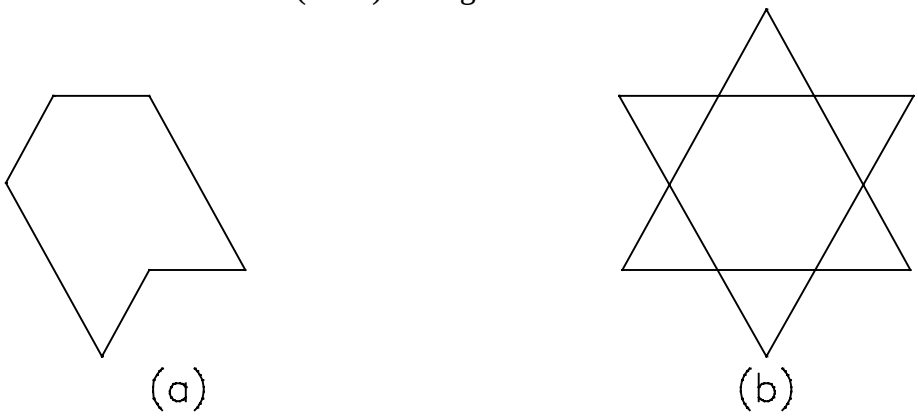
Het hierboven beschreven valideringsonderzoek levert ook de sleutel om leesbaarheid en leesvaardigheid op eenzelfde schaal te brengen. Voor een tekst T uit de teksttoets die bestaat uit zes items kunnen we voor een willekeurige waarde van θ de verwachte gestandaardiseerde score berekenen met de formule

$$\mathscr{E}(X_T) = \frac{\sum_{i \in T} a_i f_i(\theta)}{\sum_{i \in T} a_i}. \qquad (7.8)$$

Stellen we nu dat beheersing van de tekst gelijk staat met een gestandaardiseerde verwachte score van minstens *c* (bijvoorbeeld 0.7), dan kan in het rechterlid van (7.8) θ zo bepaald worden dat de verwachte score gelijk is aan *c*. We duiden deze waarde aan als θ_c. Uit de zeer hoge correlatie tussen de gemiddelde $\hat{\pi}_i$-waarden en de CLIB volgt dat de CLIB-waarde voor deze tekst in de populatie van personen met $\theta = \theta_c$ ongeveer gelijk zal zijn aan *c*. Omgekeerd -en in de mate dat het verband tussen CLIB en leesvaardigheidstoets te veralgemenen is- volgt dat een tekst met CLIB-waarde gelijk aan *c*, begrepen wordt door personen met een θ-waarde groter θ_c. Kennen we de θ-waarde van een persoon en de CLIB-waarde van een tekst, dan hebben we een rationele grond om te beslissen of de tekst al dan niet voor die persoon geschikt is. Omdat θ geschat moet worden, wordt de schatting natuurlijk niet gebaseerd op één tekst met zes items, maar op een teksttoets van redelijke lengte, zodat de meetfout (dit is de schattingsfout van $\hat{\theta}$) voldoende klein wordt gehouden.

7.3 De diagnostische verborgen-figurentest

Binnen de cognitieve psychologie worden trainingsprogramma's opgesteld om het cognitieve functioneren te beïnvloeden en om eventuele achterstanden weg te werken. Het 'Instrumental Enrichment'-programma van Feuerstein (1980) neemt hier een leidende positie in. Het programma bestaat uit 14 instrumenten die voornamelijk oefeningen in de vorm van testfiguren bevatten. Het is de bedoeling om via deze training de cognitieve capaciteiten en het algemene leervermogen van adolescenten te verhogen. Een van de instrumenten die Feuerstein gebruikte om zijn programma te evalueren is de verborgen-figurentest (Embedded Figures Test, verder afgekort als EFT), ontwikkeld door Witkin (1950). In figuur 7.7 is een item uit deze test afgebeeld.



Figuur 7.7
Voorbeeld van een verborgen-figuren opgave

De eenvoudige figuur (*a*) zit verborgen in het complexe patroon (*b*). Bij toepassing van Witkins test wordt aan de persoon eerst gevraagd het complexe patroon te beschrijven; daarna moet de eenvoudige figuur gememoriseerd worden, en tenslotte moet aangewezen worden waar de eenvoudige figuur in het complexe patroon verborgen zit. De antwoordtijd en de correctheid van het antwoord worden genoteerd.

Uit de evaluatiestudie bleek dat de personen die het 'Instrumental Enrichment' programma hadden gevolgd, gemiddeld sneller antwoordden en meer juiste antwoorden gaven dan een controlegroep die een minder specifiek trainingsprogramma had gevolgd. Bradley (1983) betoogde echter dat uit dit resultaat niet volgt dat door het trainingsprogramma cognitieve strategieën gewijzigd kunnen worden. Immers, uit de

verschillen in antwoordtijd en aantal items juist volgt niet automatisch dat er andere cognitieve strategieën gebruikt worden in de twee condities. Het probleem met de interpretatie van de EFT wordt bijvoorbeeld duidelijk geïllustreerd door de vele theoretische interpretaties die Witkin zelf en anderen aan de test hebben gegeven (Witkin & Goodenough, 1981; Pennings, 1991). In meer algemene termen gesteld, betekent dit dus dat er problemen zijn met de constructvaliditeit van de EFT. Het is niet zonder meer duidelijk wat de EFT eigenlijk meet. Op basis van een theoretische studie over de gebruikte strategieën in de EFT, kwam Pennings (1988) tot de volgende conclusies:

(1) Zeer korte antwoordtijden komen tot stand door het gebruiken van een simultane (ook genoemd holistische, synthetische of figuratieve) strategie, waarbij vorm, grootte en positie van de eenvoudige figuur als geheel in gedachten worden gehouden bij het bekijken van het complexe patroon. Het antwoord komt tot stand door een 'matching' van deze voorstelling met een gedeelte van het complexe patroon;

(2) middellange antwoordtijden resulteren bij gebruik van een successieve (analytische) strategie, waarbij onderdelen van de eenvoudige figuur (bijvoorbeeld een lijnstuk) successievelijk opgezocht worden in het complexe patroon;

(3) als de antwoordtijden, bij volwassenen en adolescenten, heel lang worden, kan toch een oplossing gevonden worden door het externaliseren van oplossingsoperaties, zoals het volgen van bepaalde lijnstukken met een aanwijsstokje op het complexe patroon;

(4) wanneer kinderen de items erg moeilijk vinden, vinden ze toch vaak de oplossing als ze een doorzichtig figuurtje in de vorm van de eenvoudige figuur mogen manipuleren over het complexe patroon. Dit wordt aangeduid als een globaal-manipulatorische strategie.

Deze vier genoemde strategieën komen bovendien overeen met een ontwikkelingslijn in de cognitieve ontwikkeling van kinderen: van een globaal-manipulatorische strategie, die helemaal extern is, naar een geïnternaliseerde strategie die verloopt van successieve en gecontroleerde operaties naar simultaan en geautomatiseerd. De vier beschreven strategieën in de volgorde (4) tot (1) weerspiegelen dus ook de chronologische ontwikkeling in het normale functioneren van een kind.

Om deze strategieën meer zichtbaar te maken dan door de pure tijdopname in de EFT, ontwikkelde Pennings een variant, het Verborgen-Figuren Diagnosticum genaamd. Daarbij wordt eenzelfde soort items gebruikt als in de EFT, doch de wijze van afname en de scoring is verschillend. De algemene procedure is een 'antwoord-totdat-juist' procedure:

(1) een juist antwoord binnen vijf seconden wordt geïnterpreteerd als evidentie voor een (succesvolle) simultane strategie, en levert een score op van vier punten;

(2) bij geen of een fout antwoord onder conditie (1), krijgt de proefpersoon speciale instructie om een successieve strategie te gebruiken. Een juist antwoord binnen de 55 seconden levert drie punten op;

(3) indien (2) niet succesvol is, krijgt de proefpersoon staafjes die in lengte overeenkomen met de lijnstukken van de eenvoudige figuur, die op het complexe patroon kunnen worden neergelegd om de eenvoudige figuur te vormen (maximale tijd 75 seconden). Succes levert een score van twee punten op;

(4) indien nog steeds geen oplossing is gevonden, kan de proefpersoon manipuleren met een doorzichtig perspex model van de eenvoudige figuur (maximale tijd 45 seconden). Een goed antwoord levert één punt op. Lukt het niet binnen de maximaal toegestane tijd dan is de itemscore nul punten.

De belangrijkste vraag met betrekking tot de constructvaliditeit van het aldus geconstrueerde meetinstrument is of deze scoringsregel zinvol is: bestaat er een abstract unidimensionaal begrip θ, zodat een grotere waarde van θ een hogere verwachte score betekent op elk item in de test. Een geschikt model om deze vraag te beantwoorden is OPLM voor polytome data (zie hoofdstuk 5).

De data waren afkomstig van 480 kinderen, 30 jongens en 30 meisjes in de leeftijd van 5, 6, 7, 8, 9, 10, 11 en 12 jaar. De test bevat zes items en de resultaten van de CML-schattings- en toetsingsprocedure zijn weergegeven in tabel 7.3. Hoewel de passing van het model niet overweldigend is, is er ook geen duidelijke evidentie om het model te verwerpen. De conclusie dat de scoringsregel zinvol is, wordt door deze analyse dus goeddeels gesteund.

Het tweede aspect van de hypothese, namelijk dat θ de individuele ontwikkeling weerspiegelt, kan gevalideerd worden door het verband tussen de leeftijd van de proefpersonen en θ te onderzoeken. Op dezelfde wijze als in paragraaf 7.1 wordt een lineair model gespecificeerd voor de latente variabele θ:

$$\theta_{vjk} = \mu + \alpha_j + \beta_k + \varepsilon_{vjk} \qquad (7.9)$$

waarin het residu ε_{vjk} normaal verdeeld is met gemiddelde nul en gemeenschappelijke variantie σ^2. Hoewel leeftijd een continue variabele is, werd de totale groep opgesplitst in vier leeftijdscategorieën: 1 = 5-6 jaar; 2 = 7-8 jaar; 3 = 9-10 jaar en 4 = 11-12 jaar.

Tabel 7.3

Parameterschattingen en toetsen voor de diagnostische EFT

Item	Cat.	*a*	β	*SE*(β)	*S*	*vg*	*p*	*M*	*M*2	*M*3

1	1	4	-.931	.085	---	-	---	3.17	-.09	-.30
	2		-.275	.046	1.41	3	.702	1.44	-.02	.26
	3		-.104	.035	5.70	4	.222	-2.12	-1.27	-1.89
	4		.582	.040	2.54	3	.467	-1.37	-1.91	-.86
2	1	3	-.815	.093	---	-	---	-1.49	-.30	-.68
	2		-.459	.060	7.38	3	.061	-1.49	.03	.03
	3		-.035	.045	1.65	5	.895	-.87	-.36	-.95
	4		.317	.044	13.06	5	.023	.01	1.41	-.68
3	1	2	-.398	.100	.42	3	.937	.72	.30	.22
	2		-.336	.082	4.74	5	.448	.61	2.02	1.74
	3		.149	.072	8.41	6	.209	1.28	.49	1.80
	4		.271	.074	3.39	5	.640	-1.51	-.97	-1.56
4	1	3	-.697	.073	.12	1	.730	.98	-.66	-.62
	2		-.126	.054	9.01	4	.061	2.70	2.44	2.84
	3		-.130	.045	3.70	5	.594	.14	-.05	.56
	4		.797	.057	1.28	3	.734	.37	.00	.86
5	1	3	-.507	.053	4.32	3	.229	-2.22	-.12	-.90
	2		.147	.043	2.91	5	.714	.72	.63	.85
	3		.407	.050	9.46	4	.051	.40	2.56	1.65
	4		1.082	.108	---	-	---	-.86	4.52	2.63
6	1	4	-.288	.043	1.25	3	.742	-.07	-.63	.89
	2		-.009	.037	4.35	4	.361	-2.46	-1.51	-2.43
	3		.344	.037	3.79	4	.435	-.21	-.58	.00
	4		1.016	.088	---	-	---	.01	-.21	-.57
R_{1c} = 85.80 (*vg* = 67; *p* =.061)										

De effecten van de leeftijdscategorieën worden weergegeven door de parameters β_k. Omdat Witkin ook verschillen tussen jongens en meisjes rapporteert voor de EFT, werd geslacht als tweede achtergrondvariabele meegenomen. De effectparameters zijn α_j (1 = 'jongen', 2 = 'meisje'). De resultaten zijn weergegeven in tabel 7.4. De schaal waarop de resultaten zijn gerapporteerd is zo geconstrueerd dat de som van de categorieparameters gelijk is aan 0 en het produkt van de discriminatie-indices gelijk is aan 1. De analysemethode is identiek aan de methode beschreven in paragraaf 7.1.

Tabel 7.4

Effectschattingen van het onderzoek met de diagnostische EFT

Parameter	Schatting	Stand. fout (*SE*)	z=schatting/*SE*
σ^2	0.54		
μ	-1.50	0.14	-10.77
α_1	0	---	---
α_2	0.12	0.09	1.26
β_1	0	---	---
β_2	1.56	0.14	10.96
β_3	2.00	0.14	14.00
β_4	2.62	0.14	18.12

De binnengroeps-standaardafwijking, σ, is gelijk aan $\sqrt{.54} = 0.735$. Het verschil tussen de tweede leeftijdsgroep en de referentiegroep (de jongste kinderen), $\beta_2 - \beta_1$, bedraagt dus meer dan twee maal de binnengroeps-standaardafwijking, terwijl de verandering van de tweede naar de volgende leeftijdsgroepen veel minder sterk uitgesproken is. De resultaten van deze analyse bevestigen dus zeer duidelijk de hypothese dat θ de individuele ontwikkeling weerspiegelt.