6

Itemresponstheorie en onvolledige gegevens

In onderzoek in de gedragswetenschappen komt het veelvuldig voor dat men niet alle gegevens bij alle personen die aan een onderzoek meedoen heeft kunnen of willen verzamelen. Onderzoek waarbij de itemresponstheorie (IRT) wordt toegepast, vormt hierop geen uitzondering. Het ontbreken van gegevens of data kunnen we ons in deze situatie als volgt voorstellen. Als we de antwoorden van personen op items of vragen weergeven in een datamatrix en als we aannemen dat in totaal n personen en k items in het onderzoek betrokken zijn, dan zal een aantal van de in totaal $n \times k$ cellen van deze matrix leeg zijn. De lege cellen vertegenwoordigen de ontbrekende gegevens of 'missing data' in het onderzoek. De redenen voor het ontbreken van gegevens kunnen van onderzoek tot onderzoek sterk variëren maar zijn globaal in te delen in drie categorieën. Het criterium voor deze indeling is de mate waarin de onderzoeker zelf het optreden van de ontbrekende gegevens onder controle heeft. De eerste categorie die we onderscheiden is dat de onderzoeker van te voren vastlegt aan welke (groep) respondenten welke items worden voorgelegd en van te voren dus ook weet waar de lege cellen in de matrix zullen zitten. Een voorbeeld hiervan is dat bij een enquête de getrokken steekproef van respondenten vanwege de lengte van de vragenlijst beurtelings het eerste deel, met algemene vragen, en het tweede deel van een vragenlijst wordt voorgelegd, dan wel het eerste en het derde en laatste deel van de lijst. De tweede categorie is dat de onderzoeker vastgelegd heeft volgens welke procedure lege cellen in de datamatrix kunnen ontstaan, maar van te voren niet exact kan voorspellen waar de cellen precies leeg zullen zijn. In het hetzelfde voorbeeld van een enquête zou dit het geval zijn als we niet beurtelings, maar op grond van de uitkomst van een worp met een munt of bijvoor-beeld op grond van de leeftijd van de respondent zouden bepalen wie welk deel van de vragenlijst gaat beantwoorden. De derde en laatste categorie van het optreden van ont-brekende gegevens is dat zonder dat de onderzoeker daar enige invloed op heeft gegevens ontbreken. Bij een enquête is dit bijvoorbeeld het geval als een respondent weigert op een bepaalde vraag antwoord te geven.

De eerste twee categorieën van ontbrekende gegevens noemt men wel structureel onvolledig, de laatste categorie ontstaat spontaan tijdens het waarnemen en zijn vanuit het gezichtspunt van de onderzoeker doorgaans ongewenst en storend. Bij de laatste categorie kan de analyse van de gegevens vaak alleen maar goed plaatsvinden als we aannames doen omtrent de mechanismen die de ontbrekende gegevens veroorzaken. Meestal zijn deze aannames niet of heel moeilijk toetsbaar. Met structureel onvolledige gegevens kennen we deze mechanismen en kunnen we in de analyse doorgaans veel beter uit de voeten. In dit hoofdstuk zullen wij ons bezighouden met structureel onvolledige designs. In de itemresponstheorie wordt namelijk met modellen gewerkt die onder bepaalde voorwaarden erg goed structureel onvolledige gegevens kunnen analyseren. Ook de niet structureel ontbrekende gegevens komen in de psychometrische praktijk voor. Denk hierbij aan ontbrekende gegevens die ontstaan doordat leerlingen opgaven in een toets overslaan of ook wel de situatie waarin de toets een zodanige lengte heeft dat sommige leerlingen bepaalde opgaven niet bereiken. We zullen deze onder-werpen niet bespreken. Voor voorbeelden van modellen die rekening houden met een tijdslimiet op de toetsafname verwijzen we naar Verhelst, Verstralen en Jansen (1993).

In het hiernavolgende zullen we eerst de relatie tussen IRT en onvolledige gegevens in het algemeen bespreken. Daarna wordt een overzicht gegeven van de in de praktijk veel voor-komende designs. In paragraaf 6.2 doen we dit door middel van het beschrijven van de datamatrices in onvolledige designs. In paragraaf 6.3 gebeurt dit aan de hand van het stochastische mechanisme dat de onvolledige gegevens veroorzaakt. Wij bespreken daarbij de drie in de praktijk meest gebruikte stochastische designtypen. Als we IRT toepassen beginnen we met het calibratie-onderzoek, het schatten van de itemparameters. Daarom zullen we hierna uitvoerig ingaan op de mogelijkheden en voorwaarden voor calibratie in onvolledige designs. Beide schattingsmethoden uit hoofdstuk 4, met behulp van de marginale aannemelijkheidsfunctie (MML) en met behulp van de conditionele aannemelijkheidsfunctie (CML) worden behandeld. In paragraaf 6.4 bespreken we de algemene voorwaarden, terwijl in 6.5 uitgebreid de mogelijkheden in de stochastische designs aan de orde komen. In paragraaf 6.6. zullen we tenslotte nog kort ingaan op het schatten van persoonsparameters in onvolledige designs.

6.1 De relatie tussen onvolledige gegevens en IRT

Alhoewel de itemresponstheorie in het algemeen een aantal voordelen heeft boven de klassieke testtheorie (zie hoofdstuk 4), komen deze voordelen vooral goed tot uitdrukking als we IRT gaan toepassen in problemen waarbij er sprake is van onvolledige gegevens. Anderzijds is het zo, dat veel van de specifieke toepassingen van IRT alleen maar mogelijk zijn omdat onvolledige gegevens analyseerbaar zijn. In zekere zin is het dus zo dat IRT en onvolledige gegevens elkaar nodig hebben. Wij gaan hier aan de hand van enkele voorbeelden nader op in.

Een veel genoemde en geroemde eigenschap van IRT is dat personen met verschillende opgaven op dezelfde schaal gemeten kunnen worden. Ofwel iets nauwkeuriger geformuleerd, indien het IRT-model geldt voor een verzameling items in een of andere goed gedefinieerde populatie dan is het mogelijk de vaardigheid van personen uit deze populatie te schatten op dezelfde schaal op basis van antwoorden van verschillende deelverzamelingen items. Deze eigenschap maakt het bijvoorbeeld mogelijk om van twee verschillende toetsen met ver-schillende opgaven de resultaten op dezelfde schaal te vergelijken. Als de itemparameters van de items bekend verondersteld kunnen worden, dan kunnen we nagaan of verschillen in prestaties tussen bijvoorbeeld jaargroepen echte verschillen zijn zonder dezelfde opgaven te laten maken. Daarbij kunnen we een mogelijke alternatieve verklaring voor verschillen tussen groepen, dat de opgaven qua moeilijkheid verschillen, zoals die onder het klassieke testmodel mogelijk is, uitsluiten. Op de mogelijkheden en technieken om deze zogenaamde geëquiva-leerde toetsen te verkrijgen wordt in hoofdstuk 9 uitvoerig ingegaan. Hier wordt het slechts als voorbeeld genoemd van een toepassing van IRT die de analyse van een onvolledige data-matrix nodig heeft: twee groepen personen maken elk slechts een deel van de totale verzameling opgaven.

Een tweede algemeen genoemd voordeel van IRT is dat de itemparameters van IRT-modellen in meer of in mindere mate onafhankelijk van de getrokken steekproef geschat kunnen worden. Indien conditionele schattingsmethoden voor de itemparameters toepasbaar zijn, zoals in het Raschmodel en in het OPLM model (zie hoofdstuk 4 en 5), behoeven er zelfs in het geheel geen aannames te worden gedaan omtrent de verdeling van de vaardigheid van de steekproef waarmee de itemparameters geschat worden. Van deze eigenschap maken we natuurlijk gebruik als we van grotere verzamelingen items de parameterwaarden op dezelfde schaal willen hebben. Dit zogenaamde calibreren van de items gebeurt vaak op basis van gegevens uit onvolledige designs. Met name is dit het geval als we itembanken, hoofdstuk 1, gaan opbouwen met gecalibreerde opgaven. Het is in calibratie-onderzoek vaak alleen al praktisch

onmogelijk, vanwege de beschikbare testtijd, om alle opgaven aan alle leerlingen in de steekproef voor te leggen. Vanwege de genoemde eigenschap van de steekproef-onafhankelijkheid van de itemparameterschattingen is dit in IRT-modellen ook niet nodig.

6.1.1 Efficiëntie van de schattingen

Zijn er enerzijds vaak praktische redenen aanwezig die noodzaken tot onvolledige designs, in toepassingen van IRT zijn het doorgaans overwegingen van efficiëntie die leiden tot het gebruik van onvolledige designs. Met efficiëntie wordt hier bedoeld de statistische efficiëntie van de schattingen van de parameters.

We zullen aan de hand van een voorbeeld illustreren, dat de standaardfout van de itemparameterschattingen kleiner is naarmate de vaardigheid van de steekproef op basis waarvan de parameters worden geschat meer overeenkomt met de moeilijkheid van de items. In dit voorbeeld gebruiken we drie gesimuleerde dataverzamelingen. Deze dataverzamelingen hebben gemeenschappelijk dat ze elk uit 1000 antwoorden op 10 items bestaan. Verder is gemeenschappelijk dat alle items in elke dataverzameling het Raschmodel volgen:

$$P(X_i = 1 \mid \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}. \qquad (6.1)$$

Dat wil zeggen: item i, met moeilijkheid β_i, wordt door een persoon met vaardigheid θ, met de in (6.1) gegeven kans goed ($\{X_i = 1\}$) gemaakt. Tenslotte is gemeenschappelijk dat bij elke dataverzameling de vaardigheid van de personen aselect getrokken wordt uit de normale verdeling met gemiddelde 0 en variantie 1: θ is $N(0,1)$ verdeeld. De drie simulaties onderscheiden zich doordat de itemmoeilijkheden, waarmee de antwoorden volgens model (6.1) gegenereerd werden, verschilden. In de eerste simulatie was $\beta = 0$, in de tweede $\beta = 1$ en in de laatste $\beta = 2$. Dus steeds waren alle items in een simulatie even moeilijk, maar in de achtereenvolgende nam de moeilijkheid telkens met 1 toe en daarmee nam de overeen-stemming tussen de gemiddelde vaardigheid (0) en de itemmoeilijkheden per simulatie af.

Tabel 6.1

Geschatte itemmoeilijkheden, standaardfouten, p-waarden gesimuleerde gegevens, waarbij de afstand tussen het gemiddelde van de vaardigheid en de moeilijkheid toeneemt per simulatie

	simulatie 1			simulatie 2			simulatie 3		
item	$\hat{\beta}$	$SE(\hat{\beta})$	p	$\hat{\beta}$	$SE(\hat{\beta})$	p	$\hat{\beta}$	$SE(\hat{\beta})$	p
1	-0.120	.066	.528	0.072	.072	.281	-0.051	.086	.166
2	-0.076	.066	.519	-0.108	.070	.313	0.059	.088	.153
3	0.056	.066	.492	-0.080	.070	.308	0.016	.087	.158
4	-0.022	.066	.508	0.060	.072	.283	-0.170	.084	.181
5	-0.018	.066	.507	-0.047	.071	.302	0.033	.088	.156
6	0.031	.066	.497	-0.019	.071	.297	-0.035	.086	.164
7	0.046	.066	.494	-0.008	.071	.295	0.024	.088	.157
8	0.002	.066	.503	0.196	.073	.260	-0.010	.087	.161
9	-0.037	.066	.511	-0.058	.071	.304	0.050	.088	.154
10	0.139	.066	.475	-0.008	.071	.295	0.085	.089	.150

Het resultaat van de itemparameterschattingen met de standaardfouten en de klassieke p-waarden van de aldus gegenereerde antwoorden, bepaald met het programma OPLM (Verhelst, Glas & Verstralen, 1993), staan in tabel 6.1.We zien duidelijk het effect, dat de standaardfouten van de itemparameters kleiner zijn naarmate de vaardigheid van de steekproef beter in overeenstemming is met de moeilijkheid van de items, hoewel het aantal waarnemingen voor alle items 1000 is. De itemmoeilijkheden in de eerste simulatie worden het nauwkeurigste geschat. Naarmate de gemiddelde vaardigheid verder afligt van de moeilijkheid van de items wordt de standaardfout groter. Opgemerkt kan nog worden dat de standaardfouten van de items per simulatie ook enigszins verschillen, hetgeen veroorzaakt wordt doordat ook de SE's geschat worden (zie hoofdstuk 4).

Dit eenvoudige voorbeeld moge duidelijk maken dat de efficiëntie van de itemparameter-schattingen in het algemeen verhoogd kan worden door moeilijkheid en vaardigheid op elkaar af te stemmen. De efficiëntie van statistische schattingen wordt doorgaans uitgedrukt in het verschil of in de verhoudingen tussen de zogenaamde statistische informatie (zie hoofdstuk 4) die in een gegevensverzameling met betrekking tot een parameter aanwezig is. Voor een kwantificering van de informatiewinst met betrekking tot de itemparameterschattingen bij bepaalde onvolledige designs verwijzen wij naar Verhelst (1989). Het zal duidelijk zijn dat principieel dezelfde argumentatie geldt voor de schatting van de persoonsparameters en of van de kenmerken van de populatie personen: deze schattingen zullen efficiënter zijn naarmate de moeilijkheid van de voorgelegde items beter is afgestemd op de vaardigheid. In praktijk-toepassingen zijn, in tegenstelling tot het hiervoor geschetste voorbeeld, de items niet even moeilijk en hebben de personen niet dezelfde vaardigheid. We kunnen dus aan efficiëntie

winnen door de moeilijkste items aan de meest vaardige personen voor te leggen en de gemakkelijkste aan de minst vaardige. Dit resulteert uiteraard in een onvolledig design.

6.1.2 Calibratie in onvolledige designs en linken

Met name in de Amerikaanse psychometrische literatuur, bijvoorbeeld Hambleton en Swaminathan (1985), wordt calibreren in onvolledige designs vaak beschreven als een activiteit die in twee fasen uiteenvalt. De eerste is het calibreren in volledige deeldesigns, waarna in de tweede fase de parameters, om onderling vergelijkbaar te kunnen zijn, via het zogenaamde 'linken' op dezelfde schaal worden gebracht. Men noemt dit ook wel het equivaleren van de itemparameters.

Zoals bekend (hoofdstuk 4) wordt tijdens het calibratieproces de schaal op enigszins arbitraire wijze gefixeerd. We fixeren de schaal tijdens de calibratie, als we met de CML-schattingsmethode werken, zoals in het Raschmodel en het OPLM model vaak door de som van de geschatte itemmoeilijkheden (en dus ook het gemiddelde) op 0 te stellen: $\sum_{i=1}^{k} \hat{\beta}_i = 0$. Een andere mogelijkheid die veelal wordt toegepast bij calibratie met MML is de schaal te fixeren zodanig dat het gemiddelde van de steekproefverdeling van de vaardigheid θ vastgelegd wordt op 0 en de variantie van deze verdeling op 1. In het algemeen is het echter zo dat we de gekozen schaal op willekeurige wijze lineair kunnen transformeren. Zoals uiteengezet in hoofdstuk 4 veranderen we daardoor slechts het willekeurig te kiezen nulpunt en de eenheid van de schaal.

Als voorbeeld hiervan blikken we even terug op de resultaten van tabel 6.1 Daar zien we dat de geschatte moeilijkheden tussen de simulaties nauwelijks verschillen, ondanks dat we weten dat er wel verschillen zijn. Duidelijk is dat te zien in tabel 6.1 aan de klassieke p-waarden. Waaruit volgt dat per calibratie de schaal op dezelfde willekeurige wijze gefixeerd is en dat de waarden van de itemparameters per simulatie op een andere niet vergelijkbare schaal liggen. Om de moeilijkheidsschattingen van de items in de drie simulaties te kunnen vergelijken zullen er nog transformaties nodig zijn die de parameterschattingen op dezelfde schaal brengen.

Hoe dit in zijn werk zou kunnen gaan, zullen we toelichten met een ander voorbeeld. In dit voorbeeld hebben we een onvolledig design en wordt in twee aparte calibraties de schaal gefixeerd, waarna er bij het verbinden van de schalen ervoor gezorgd wordt dat de itemparameters van beide groepen items op dezelfde schaal komen te liggen. Dit komt neer op het vinden van een transformatie van een van de, of eventueel van beide, gecalibreerde schalen. Zo'n transformatie kan op verschillende manieren worden

bepaald en uitgevoerd. Een ervan zullen we met ons voorbeeld toelichten. We beschouwen een design met twee groepen van tien items en twee groepen personen. Hierbij zijn item 1 tot en met 5 gemaakt door de eerste groep, de items 6 tot en met 10 alleen door tweede en de items 11 tot en met 15 door beide groepen. Om zeker te zijn de items aan een IRT-model voldoen, zijn antwoorden op de items conform het Raschmodel (6.1) gegenereerd. In beide groepen werden 1000 antwoordpatronen gegenereerd. De calibratie van de items in beide groepen apart, dat wil zeggen per volledig deeldesign, met de CML-schattingsmethode van het programma OPLM leverde de in tabel 6.2 gegeven schattingen van de moeilijkheid op.

We zien in tabel 6.2 dat voor item 11 tot en met 15 ondanks dat het dezelfde items zijn en ondanks dat we weten zeker weten dat het Raschmodel geldt de geschatte moeilijkheden tussen de calibraties nogal verschillen. Deze verschillen kunnen twee oorzaken hebben. Kleinere fluctuaties kunnen veroorzaakt worden door de steekproef, want de steekproeven zijn eindig. Systematische verschillen worden echter veroorzaakt doordat in beide calibraties op een arbitraire wijze het nulpunt van de schaal is vastgelegd, zodanig dat de gemiddelde moeilijkheid in de te calibreren toets 0 is. De eenheid van de schaal is in dit voorbeeld van het Raschmodel op dezelfde wijze vastgelegd: alle discriminatie-indices zijn in beide calibraties gelijk aan 1 gekozen. Een manier, zie bijvoorbeeld ook Wright en Stone (1979), om alle itemparameters vergelijkbaar en dus op één schaal te krijgen is de volgende.

Tabel 6.2

Geschatte itemmoeilijkheden in een onvolledig design met overlappende items per volledig deeldesign met de verschillen tussen de gemeenschappelijke items

Item	Calibratie 1 $\hat{\beta}^{(1)}$	Calibratie 2 $\hat{\beta}^{(2)}$	$\hat{\beta}^{(2)} - \hat{\beta}^{(1)}$
1	-2.041		
2	-0.927		
3	0.093		
4	0.976		
5	1.919		
6		-0.533	
7		-0.489	
8		-0.445	
9		-0.430	
10		-0.626	
11	0.026	0.481	.455
12	-0.051	0.545	.596

13	-0.109	0.453	.562
14	0.035	0.527	.492
15	0.079	0.516	.437
Gem.	0.000	0.000	.508

Bepaal in eerste instantie de verschillen tussen moeilijkheidsschattingen van de gemeenschappelijk items. Het resultaat staat in de vierde kolom van tabel 6.2. Het gemiddelde verschil per item in geschatte moeilijkheid tussen beide calibraties is 2.542/5 = .508. Een manier om de itemparameters van de eerste calibratie op de schaal van tweede calibratie te krijgen is simpel het optellen van dit gemiddelde verschil bij alle geschatte moeilijkheden van de eerste calibratie. Het resultaat staat in tabel 6.3. Omdat we nu voor de gemeenschappelijke items 11 tot en met 15 beschikken over twee schattingen van de moeilijkheid, die variëren door statistische variatie, zouden we als uiteindelijk schattingen voor deze items het gemiddelde kunnen nemen. Het resultaat van de op deze wijze op dezelfde schaal gebrachte schattingen van de itemparameters staat in de vierde kolom van tabel 6.3. We zien dat het gemiddelde van de geschatte moeilijkheden op deze schaal 2.560/15 = .171 bedraagt.

Tabel 6.3

Het op dezelfde schaal brengen van in volledige deeldesigns geschatte itemmoeilijkheden het resultaat van een simulatie calibratie

Item	Calibratie 1 $\hat{\beta}^{(1)} + .508$	Calibratie 2 $\hat{\beta}^{(2)}$	Calibratie	Calibratie gem .00	Calibratie simultaan
1	-1.533		-1.533	-1.704	-1.703
2	-0.418		-0.418	-0.589	-0.589
3	0.601		0.601	0.430	0.431
4	1.484		1.484	1.313	1.314
5	2.427		2.427	2.256	2.256
6		-0.533	-0.533	-0.704	-0.704
7		-0.489	-0.489	-0.660	-0.660
8		-0.445	-0.445	-0.616	-0.616
9		-0.430	-0.430	-0.601	-0.601
10		-0.626	-0.626	-0.797	-0.797
11	0.534	0.481	0.508	0.337	0.339
12	0.457	0.545	0.501	0.330	0.326
13	0.399	0.453	0.426	0.255	0.253
14	0.543	0.527	0.535	0.364	0.366
15	0.587	0.516	0.552	0.381	0.384
Gem.	0.508	0.000	0.171	0.000	0.000

Daarmee hebben we dus bereikt dat de moeilijkheidsparameters van alle items op dezelfde schaal zijn gebracht en daardoor onderling vergelijkbaar zijn. Tenslotte kunnen we voor de totale itemverzameling op gebruikelijke wijze de schaal fixeren, zodanig dat gemiddelde moeilijkheid over alle items 0.000 wordt. Dit bereiken we eenvoudig door van alle geschatte moeilijkheden 0.171 af te trekken. Het resultaat staat in de vijfde kolom van tabel 6.3.

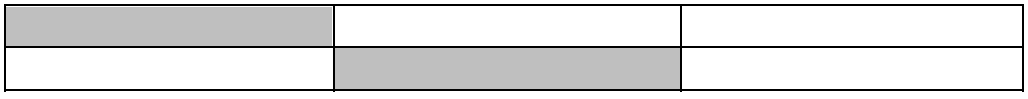
Wij zullen niet nader op ingaan op de verschillende andere manieren, die in de psycho-metrische literatuur zijn voorgesteld om in verschillende onvolledige designs een 'linktransformatie' te bepalen om parameters op één schaal te brengen. De reden hiervoor is dat het calibreren in een onvolledige gegevensverzameling ook beschouwd kan worden als een simultaan proces, waarin naast het schatten van de parameters deze tevens op dezelfde schaal worden afgebeeld. Het onderscheid in fasen, calibreren in volledige deeldesigns en vervolgens linken, dat in de literatuur vaak wordt gemaakt, is historisch ontstaan en is eigenlijk niet meer functioneel. De schattings- en toetsingstheorie voor IRT-modellen is in eerste instantie ontwikkeld voor volledige designs. En oudere computerprogrammatuur voor de calibratie kon dan ook alleen

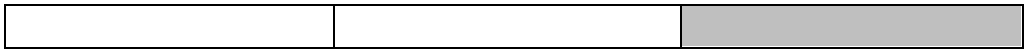
maar volledige designs analyseren en daarom moest het proces in twee fasen verlopen. Tegenwoordig is echter de theorie voor het schatten en toetsen in onvolledige designs zo ver ontwikkeld, dat ze geïmplementeerd is in programmatuur (bijvoorbeeld OPLM) zodat de traditionele omweg niet meer noodzakelijk is: calibratie vindt plaats in onvolledige designs, waarbij de itemparameters op dezelfde schaal komen te liggen door gebruik te maken van de gemeenschappelijk elementen in de deeldesigns, en de schaal wordt in een keer voor de totale gegevensverzameling gefixeerd. Ter illustratie zijn in de laatste kolom van tabel 6.3 de resultaten van de simultane calibratie van alle opgaven in de onvolledige gegevensverzameling met OPLM opgenomen. Zoals het resultaat laat zien, is er nauwelijks sprake van verschillen in de geschatte moeilijkheden. Merk echter op dat de standaardfouten van de itemparameterschattingen bij simultane calibratie kleiner worden dan bij combinatie van afzonderlijke calibraties. Zie hiervoor Vale (1986) en Verhelst (1993). Het calibreren in volledige deeldesigns en daarna de parameters op dezelfde schaal brengen of equivaleren moet dus zo mogelijk vervangen worden door simultane calibratie in een onvolledig design.

Of we in een keer in een onvolledig design de schaal fixeren, dan wel in fasen, er zal altijd tussen de volledige deeldesigns iets gemeenschappelijks moeten zijn, dat er voor kan zorgen dat de parameters op dezelfde schaal kunnen worden gebracht. De gemeenschappelijkheid kan liggen in de personen die verschillende items maken, dan wel in de items die door personen worden gemaakt. Voor deze zogenaamde ankering zijn verschillende mogelijkheden die we in de volgende paragraaf zullen bespreken. Het anker zorgt ervoor dat er een basis is voor de vergelijking tussen verschillende calibraties, dan wel dat in een calibratie de schaal eenduidig kan worden gefixeerd.

6.2 De datamatrices van structureel onvolledige designs

In deze paragraaf zullen we de in praktijk meest voorkomende structureel onvolledige designs beschrijven. We zullen dat doen door aan te geven hoe de uiteindelijk te analyseren datamatrix eruit ziet. In de figuren waarin de designs zijn gegeven, staan steeds verticaal personen en horizontaal items. Door arcering is aangegeven welke (groepen) personen welke (groepen) items hebben gemaakt. De niet-gearceerde gebieden geven de ontbrekende gegevens. Steeds zullen we aangeven hoe de in het voorgaande genoemde ankering plaatsvindt.





Figuur 6.1
Niet-verbonden of niet-geankerd design

In figuur 6.1 is schematisch een voorbeeld datamatrix weergegeven waarbij er geen overlap is tussen de drie toetsen en evenmin overlap tussen de drie groepen personen die de toetsen maken. Dit design wordt in de praktijk veel toegepast ondanks dat dit een design is, waarbij de wijze van ankering tussen de delen niet aan de datamatrix is te zien. Als de parameters van de opgaven in deze toetsen op dezelfde schaal moeten worden gebracht, zal het duidelijk zijn dat dit niet via gemeenschappelijke items of personen kan. Dus de gegevens zullen op een speciale manier verzameld moeten zijn, dan wel zullen er extra aannames nodig zijn, omtrent de wijze waarop de gegevensverzameling tot stand is gekomen om de onderdelen aan elkaar te verbinden. Een veel gebruikte opzet hierbij is dat statistisch equivalente groepen de verschillende toetsen maken, hetgeen in de praktijk goed gerealiseerd kan worden door leerlingen die aselect zijn getrokken uit een populatie aselect toe te wijzen aan de toetsen. Dit wordt dan het ’random group design’ genoemd. Omgekeerd zou men op soortgelijke wijze kunnen veronderstellen of bewerkstelligen dat er equivalente toetsen zijn. Dit komt in de praktijk minder vaak voor.

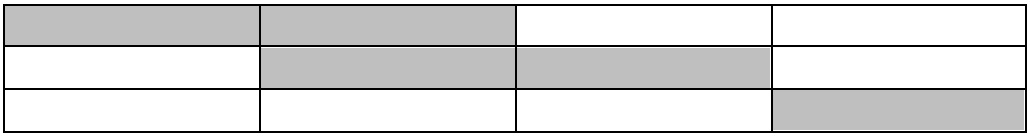
Hoe het ook zij, het niet-geankerd design, waarbij de groepen proefpersonen even groot zijn, heeft in het algemeen als voordeel dat elk item in principe even vaak wordt afgenomen. Alhoewel er geen garantie is dat een gelijk aantal afnames per item tot even nauwkeurige schattingen van de itemparameters zal leiden, geeft dit zolang er geen a priori informatie over de itemparameters of de vaardigheid van de steekproeven leerlingen aanwezig is, de beste gelegenheid voor even precieze schattingen van alle items. Bovendien is het niet-geankerde design in sommige toepassingen het enig mogelijke design. Bijvoorbeeld bij examens waarbij geheimhouding van de opgaven een belangrijke rol speelt.

Figuur 6.2
Ankeritemsdesign

Het in de praktijk traditioneel meest voorkomende design is in figuur 6.2 in zijn meest simpele vorm weergeven. In dit design met ankeritems, ’common items design’

of soms ook wel ankertoets design genoemd, wordt een deelverzameling van de items door beide onderscheiden groepen personen gemaakt. De itemparameters worden in de calibratie op een schaal gebracht via de items die gemeenschappelijk zijn afgenomen. Het zal duidelijk zijn dat dit design eenvoudig naar meer groepen items en personen kan worden gegeneraliseerd. Het belangrijkste voordeel van dit design is dat in de analyse noch de equivalentie van de groepen personen, noch van de groepen items verondersteld hoeft te worden. Een mogelijk nadeel is dat de parameters van de gemeenschappelijke items in het design nauwkeuriger geschat zullen worden dan de items die slechts in een toets voorkomen, want de gemeenschappelijke items worden door meer personen beantwoord.

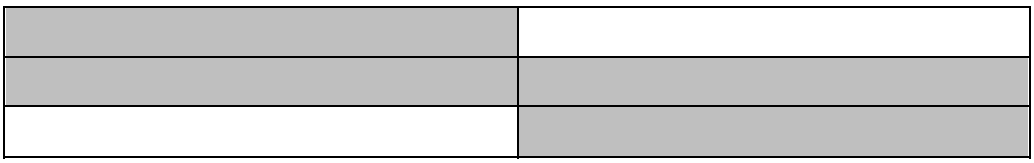
De designs die hier worden besproken komen in de praktijk om diverse redenen ook in allerlei combinaties voor. Een voorbeeld hiervan staat in figuur 6.3.



Figuur 6.3
Gedeeltelijk verbonden design

Figuur 6.3 geeft een slechts gedeeltelijk verbonden design. De items van toets 1 en toets 2 zijn via een ankertoets wel verbonden, terwijl de items van toets 3 niet verbonden zijn met de items van toets 1 of toets 2. Dit design heeft de voor- en nadelen van de basisdesigns waaruit het is samengesteld.

Een variant op het klassieke ankeritemsdesign of ankertoets design is het ankergroep-design. Zie figuur 6.4.



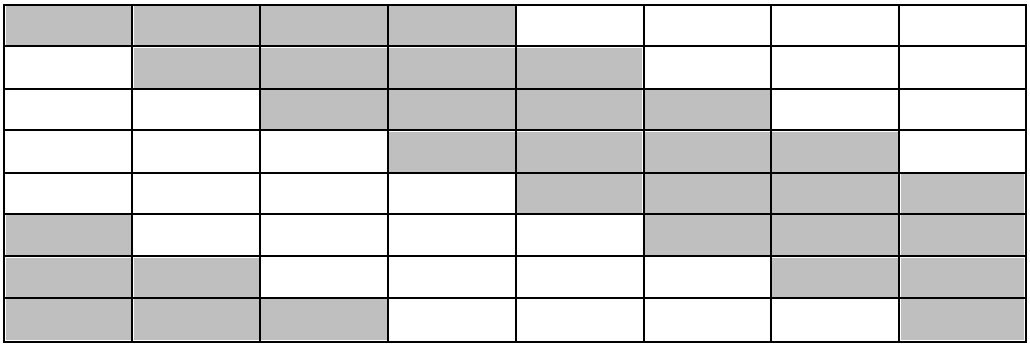
Figuur 6.4
Ankergroepdesign

Het ankergroepdesign, 'common person design,' is eigenlijk het gespiegelde van het ankeritemsdesign. De itemparameters worden op de gemeenschappelijke schaal geplaatst door de vaardigheden van de leerlingen die in dit voorbeeld de beide toetsen maken. Ook in dit design is het niet nodig aan te nemen dat groepen items of

leerlingen equivalent zijn. Alle opgaven worden in principe even nauwkeurig geschat echter ten koste van de ongelijkheid van de nauwkeurigheid waarmee personen kunnen worden geschat. Verder is een praktisch nadeel van dit design dat het moeilijk kan zijn om een groep leerlingen te vinden die alle opgaven kan maken.

Uiteraard kan men het ankergroepdesign en het ankeritemsdesign ook weer combineren en een dubbel anker leggen, zowel over personen als over groepen. Veel voordelen heeft zo'n design echter niet, men houdt namelijk het ongelijke aantal waarnemingen per item en per persoon.

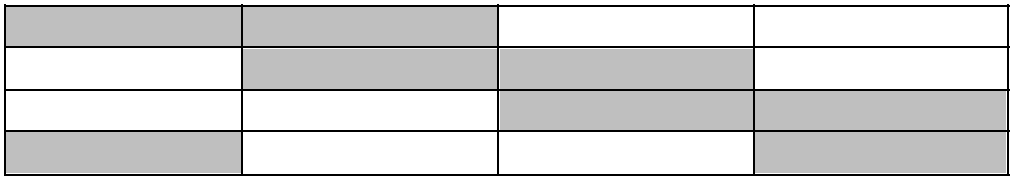
Het nadeel van het ongelijke aantal waarnemingen per opgave en per persoon wordt opgelost in zogenaamde ineengestrengelde of kettingdesigns, 'interlaced design', Vale (1986). In zijn meest extreme vorm heeft zo'n design evenveel verschillende toetsen als er opgaven of items zijn. In figuur 6.5 is een voorbeeldje gegeven met in totaal acht items waarbij elke toets bestaat uit vier items.



Figuur 6.5
Ineengestrengeld of kettingdesign: een item per blokje

De eerste toets begint met item 1 en bestaat verder uit de daaropvolgende items totdat de toets zijn vastgelegde lengte bereikt. De tweede toets begint met tweede item. Enzovoort, totdat elk item eenmaal het eerste item in een toets is geweest. Een voordeel van dit design is dat er duidelijk een ankeritem effect wordt bereikt, terwijl toch het aantal afnames per item en de toetslengte per persoon in totaal gelijk is. Indien de aldus ontstane toetsen aselect over de groepen worden verdeeld zijn ook de groepen statistisch equivalent. Het nadeel van dit design is praktisch van aard: er moeten net zoveel boekjes gedrukt als er items zijn. Dit design zal dus in toepassingen met grotere aantallen items alleen gerealiseerd kunnen worden als de items via de computer worden aangeboden. Zolang de toetsen op papier worden gedrukt is een praktische bruikbare en zeer aantrekkelijke variant van het volledig ineengestrengelde design het geblokt kettingdesign. In figuur 6.6 is daarvan een voorbeeld gegeven. De

blokken bevatten hierbij meerdere items. Als we, als in figuur 6.5, in totaal acht items hebben, bestaat elk blokje in figuur 6.6 dus uit twee items.



Figuur 6.6
Geblokt kettingdesign

In dit design zal het equivalente groepen effect wellicht minder bereikt, echter de voordelen van het design zijn evident: er zijn slechts een beperkt aantal fysieke toetsboekjes nodig en alle items worden in dit design ook weer even vaak afgenomen.

6.3 De stochastische structuur van structureel onvolledige designs

In deze paragraaf zullen we nader ingaan op de verschillende soorten structureel onvolledige gegevens design die in de IRT veel gebruikt worden. Wij onderscheiden de drie designtypen, die in de praktijk het meest voorkomen. De designs onderscheiden zich van elkaar door het mechanisme of procedure waardoor de ontbrekende gegevens in het design, lege cellen in de datamatrix, ontstaan. Dit mechanisme zullen we beschrijven als een toevalsmechanisme: door middel van kansen of verdelingen is aan te geven dat bepaalde waarnemingen wel of niet zullen voorkomen in de datamatrix. Vandaar dat we spreken over de stochastische structuur van de designs. In de paragrafen 6.5 en 6.6 zullen we bekijken in welke omstandigheden bij het schatten van de modelparameters rekening gehouden moet worden met het toevalsmechanisme dat de lege cellen in datamatrices veroorzaakt. Voor de goede orde wijzen wij erop dat bij de designs, die hierna worden beschreven, in principe alle in paragraaf 6.2 beschreven datamatrices kunnen voorkomen.

Voor de beschrijving spreken we eerst wat notatie af. In totaal beschouwen we een verzameling van k items. Hieruit worden B toetsboekjes samengesteld, geïndexeerd met $b = 1,...,B$. Elk boekje bevat $k_b, b = 1,...,B$ items, die elkaar, eventueel deels, over-lappen. Elke persoon maakt de items uit slechts één boekje. Voor elke persoon $v, v = 1,...,n$ definiëren we een zogenaamde itemindicator variabele. Deze variabele is een vector, die evenveel elementen bevat als het totaal aantal, k, opgaven: $\boldsymbol{R}_v = (R_{v1},...,R_{vk})$. Elk element van de itemindicator vector kan de waarde 1 of 0

aannemen al naar gelang de persoon het betreffende item maakt of niet. De itemindicator vector kan B verschillende waarden aannemen, net zoveel als er verschillende toetsboekjes zijn. De waarde voor bijvoorbeeld toetsboekje 1 bestaat uit een vector met een lengte van k met daarin k_1 enen en $k - k_1$ nullen op de plaatsen, die de items uit de totale verzameling indiceren, respectievelijk voor items die in het toetsboekje zitten en voor items die er niet in zitten. In het algemeen neemt de itemindicator de waarden $\boldsymbol{r}_b$ aan die staan voor een permutatie van k_b, het aantal items in toetsboekje b, enen en $k - k_b$ nullen voor, $b = 1,...,B$. Dat wil zeggen dat van een persoon v de itemindicator $\boldsymbol{R}_v$ de waarde $\boldsymbol{r}_b$ heeft, als deze persoon boekje b heeft gemaakt. In de hiernavolgende bespreking van de drie meest voorkomende stochastische designs zal steeds worden aangegeven wat de verdeling is van deze itemindicator.

6.3.1 Gerandomiseerd onvolledig design

In gerandomiseerde ofwel volledig door het toeval bepaalde designs, 'randomized-incompletedesign', besluit een onderzoeker zonder gebruik te maken van a priori kennis van de vaardigheid van de persoon met een van te voren bekende kans een van de B toetsboekjes aan een persoon toe te wijzen. In de praktijk worden in deze designs vaak uit de beschikbare itemverzameling B boekjes samengesteld, die een even groot aantal items bevatten en vaak ook nog nominaal parallel zijn, dat wil zeggen, gelijk qua inhoudelijke samenstelling en qua ingeschatte moeilijkheid. De toewijzing van een boekje aan een persoon kan natuurlijk echt aselect geschieden: elke persoon krijgt met een even grote kans, en wel $1/B$, een bepaald boekje te maken. Meer algemeen krijgt een persoon een boekje met bekende kans ϕ_b, zodanig dat $\sum_{b=1}^{B}\phi_b = 1$. In het algemeen wordt de verdeling van de itemindicator gegeven door

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b) = \phi_b. \qquad (6.2)$$

Dit geldt voor alle personen $v = 1,...,n$ en alle toetsboekjes $b = 1,...,B$. De belangrijkste reden om gerandomiseerde designs in IRT calibratie-onderzoek te gebruiken is dat het doorgaans fysiek onmogelijk is om leerlingen alle opgaven uit de verzameling te calibreren opgaven te laten maken. Zolang men bij de opzet geen gebruik kan of wil maken van a priori kennis over de vaardigheid van de leerlingen en of de moeilijkheid van de opgaven, zijn gerandomiseerde designs het meest praktisch en naar verwachting het meest efficiënt voor de calibratie van alle opgaven.

Een bijzonder geval van gerandomiseerde onvolledige designs zijn de in de praktijk vaak voorkomende a priori gefixeerde onvolledige designs. Dat zijn designs waarin de verdeling van de itemindicator gegeven wordt door

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b) = 0 \quad \text{of} \quad 1. \tag{6.3}$$

Met andere woorden, van te voren is met kans 1 bepaald wie welk toetsboekje krijgt. Van belang hierbij is op te merken, dat in de toekenning van een toetsboekje aan een persoon de kenmerken van de persoon ook geen rol spelen. Als dat het geval zou zijn dan hebben we een designtype dat in paragraaf 6.3.3 wordt besproken. Gefixeerde onvolledige designs zijn de designs die in de inleiding van dit hoofdstuk beschreven werden als de structureel onvolledige die in de eerste categorie vallen. De categorie waarbij de onderzoeker volledig onder controle heeft waar de lege cellen in datamatrix zullen zitten. De gerandomiseerde designs in het algemeen en ook de designs die hierna worden beschreven vallen onder de tweede categorie: slechts de procedure volgens welke de ontbrekende gegevens ontstaan, staat onder controle van de onderzoeker.

6.3.2 Meerfasen onvolledig design

In meerfasen designs, 'multistage testing design', is de toewijzing van items aan personen mede afhankelijk van de resultaten die de personen op een deel van de items halen. In de eerste fase krijgen bijvoorbeeld alle personen dezelfde deelverzameling items, meestal van middelmatige moeilijkheid, uit de totale itemverzameling te maken. Op grond van de scores op deze eerste groep items, die je de sorteertoets zou kunnen noemen, maken de personen in fase twee verschillende items. Bijvoorbeeld personen met hoge scores op de sorteertoets maken in fase twee een deelverzameling items uit de totale itemverzameling die van te voren wat moeilijker ingeschat wordt, terwijl personen met lage scores een verzameling gemakkelijker geachte items maken. Een simpel voorbeeld met een totale itemverzameling bestaande uit twintig items. Tien (nummers 1 tot 10) zijn er middelmatig moeilijk, vijf (item 11 tot 15) worden redelijk gemakkelijk geacht, en de laatste vijf zijn items waarvan de geschatte moeilijkheid wat hoger ligt (item 16 tot 20). Een tweefasen design zou er dan uit kunnen zien als in figuur 6.7 is aangegeven.

	Items		
Leerlingen	1 t/m 10	11 t/m 15	16 t/m 20
	$0 \leq s \leq 5$		
	$6 \leq s \leq 10$		
	Fase 1	Fase 2	

Figuur 6.7
Tweefasen design

De sorteertoets bestaat uit de middelmatig moeilijke items, is de somscore s hierop meer dan 5 dan maakt de persoon in fase twee de moeilijker ingeschatte items (16 tot 20), anders de gemakkelijker items.

Het zal duidelijk zijn dat dit sorteerproces in principe ook in een tweede fase kan worden voortgezet en in een derde fase of nog verder. Het sorteren op grond van een verzameling items hoeft natuurlijk niet plaats te vinden in twee groepen, maar evengoed kunnen meerdere groepen worden onderscheiden, die evenveel verschillende trajecten starten in de item-verzameling. Essentieel voor meerfasen toetsen is dat de selectie items die een persoon uiteindelijk maakt direct afhankelijk is van de score op items die eerder door deze persoon zijn gemaakt.

De uiteindelijke verzameling items die een persoon maakt duiden we, als eerder, weer aan met boekje b. De verdeling van de itemindicator voor een persoon in meerfasen toetsen wordt dan gegeven door de kans dat een bepaald boekje wordt gemaakt. Deze kans is 0 of 1 afhankelijk van het wel of niet voldaan zijn aan de criteria die gesteld worden aan de geobserveerde itemscores om een bepaald boekje te krijgen. In het voorbeeld uit figuur 6.7 krijgt men met kans 1 boekje 1 als $s_v = \sum_{i=1}^{10} x_{vi} \leq 5$, waarin x_{vi} de score is van persoon v op item i ,en met kans 0 boekje 2; als $s_v \geq 6$ is de kans op boekje 1 gelijk aan 0 en op boekje 2 gelijk aan 1. Algemener geldt natuurlijk ook dat als we alle itemscores van een persoon gegeven hebben, de kans op een bepaald boekje ook 0 of 1 is. Als we de vector van de van persoon v geobserveerde itemscores schrijven als $\boldsymbol{X}_{obs,v}$, met obs,v de verzameling van alle itemnummers of indexen die deze persoon maakt, dan geldt

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid \boldsymbol{x}_{obs,v}) = 0 \quad \text{of} \quad 1. \tag{6.4}$$

Dit geldt weer voor alle personen $v = 1,...,n$ en alle toetsboekjes $b = 1,...,B$.

Het idee achter meerfasen toetsen is dat daarmee de efficiëntie van de schattingen kan worden verhoogd, doordat met de toewijzing van de items aan persoon afstemming

plaats vindt tussen de van te voren ingeschatte moeilijkheid van de items en de tussentijds ingeschatte vaardigheid van de personen. Het zal duidelijk zijn dat naarmate er meer fasen worden onderscheiden in principe het afstemmen van moeilijkheid op vaardigheid nauwkeuriger kan gebeuren. Meerfasen designs vinden toepassing bij zowel calibratie-onderzoek als in situaties waarin we bijvoorbeeld met behulp van een gecalibreerde item-verzameling persoonsparameters willen schatten. Adaptief toetsen is eigenlijk een limietgeval van meerfasen toetsen; daarbij zijn er voor elke persoon evenveel fases als hij of zij items maakt. Het aantal items zal hierbij per persoon in het algemeen verschillen. Na elke itemafname wordt op grond van een voorlopige schatting van de vaardigheid, gebaseerd op de tot dan toe gemaakte items, een nieuw item gekozen waarvan de moeilijkheid het best in overeenstemming met deze vaardigheid. Gestopt wordt met toetsen, zodra de vaardigheid van de persoon met vooraf vastgestelde nauwkeurigheid kan worden geschat. Adaptief toetsen wordt in calibratie opzetten niet toegepast omdat criteria om het beste item uit een verzameling beschikbare te kiezen eigenlijk alleen met bekend (veronderstelde) itemparameters goed gekwantificeerd kunnen worden. Als het gaat om de vaardigheid van personen te schatten is adaptief toetsen de meest efficiënte vorm van toetsen.

6.3.3 Groepsgericht onvolledig design

In groepsgerichte designs, 'targeted testing design', wordt de toewijzing van de items aan de personen bepaald op basis van te voren bekende achtergrondinformatie van de persoon. Die achtergrondinformatie kunnen we uitdrukken door de waarden die een toevalsvariabele Y aanneemt. Dan hangt Y doorgaans positief samen met de vaardigheid van de leerlingen. Groepsgerichte designs zien er dan zo uit dat de gemakkelijker geachte boekje(s) gemaakt worden leerlingen met waarden van Y die naar verwachting samengaan met een geringere vaardigheid; leerlingen met Y waarden die duiden op een hogere vaardigheid maken de naar verwachting moeilijke boekje(s). Efficiëntie winst in de schatting door betere afstemming van de vaardigheden op de moeilijkheden wordt hierbij weer verwacht. Zonder dat dit de algemeenheid beperkt, nemen we aan dat we van de achtergrondvariabele Y evenveel waarden onderscheiden als verschillende toetsboekjes (B) in het design. Die waarden zijn dus in het algemeen $y_1,...,y_B$. Bij elke waarde y_b wordt een ander boekje b gemaakt. Dit boekje bestaat uit een deelverzameling items uit de totale itemverzameling. De waarde van de itemindicator van een persoon die dit boekje maakt is $\boldsymbol{r}_b$. Dan kunnen we als voorheen de verdeling van de itemindicator in groepsgerichte designs schrijven als:

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid Y_v = y_b) = 1,$$
$$P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid Y_v \neq y_b) = 0, \tag{6.5}$$

voor alle personen $v = 1,...,n$ en voor alle te onderscheiden waarden van de achtergrond-variabele $b = 1,...,B$.

Bij groepsgerichte designs zijn twee situaties te onderscheiden met betrekking tot de rol die de achtergrondvariabele in de analyse en eventueel in de steekproeftrekking speelt. In de eerste is de rol van de achtergrondvariabele zeer beperkt: hij wordt alleen maar gebruikt om de efficiëntie van de schattingen te verhogen en zijn we niet geïnteresseerd in de resultaten van leerlingen met bepaalde waarden van de achtergrondvariabele. De tweede en in de praktijk meest voorkomende rol van de achtergrondvariabele is dat we ook in de vaardigheids-verdelingen bij verschillende waarden van achtergrondvariabele geïnteresseerd zijn. De totale populatie wordt door de achtergrondvariabele opgedeeld in een aantal subpopulaties die ons interesseren.

Een concreet voorbeeld van de eerste situatie deed zich voor bij het Periodiek Peilings Onderzoek (PPON) in het basisonderwijs (Verhelst & Eggen, 1989), waarbij het geschatte niveau van de leerling door de leerkracht bepaalde welke toets de leerling maakte. Dit voorbeeld wordt uitgebreid besproken in paragraaf 7.1. Hier zij slechts vermeld dat in dit onderzoek het leerkrachtoordeel gebruikt werd om de efficiëntie van het design te verhogen, zonder dat men geïnteresseerd in de variabele zelf.

De tweede situatie komt in de praktijk regelmatig voor. Behalve in de itemparameters zijn we ook geïnteresseerd in de vaardigheidsverdelingen van de onderscheiden groepen. Stel dat we bijvoorbeeld een verzameling items die luistervaardigheid meten, willen calibreren voor de populatie van leerlingen uit het derde leerjaar van het VBO en het MAVO. In dat geval zal de verdeling van de vaardigheid in de subpopulaties VBO en MAVO zeker interessant zijn. In de praktijk komt de interesse in de verschillende vaardigheidsverdelingen daarbij vaak expliciet naar voren als men ten behoeve van het calibratie-onderzoek geen aselecte steekproef uit de totale populatie van derde klassers VBO en MAVO trekt, maar een gestratificeerde steekproef: per schooltype trekt men een aselecte steekproef. Om er zeker van te zijn dat per subpopulatie de vaardigheidsverdelingen even nauwkeurig kunnen worden geschat, zijn de aantallen leerlingen uit de subpopulaties in de steekproef vaak even groot, maar de proporties uit de verschillende subpopulaties niet noodzakelijk gelijk aan de proporties in de totale populatie. Zodat we niet meer beschikken over een aselecte steekproef uit de totale populatie.

6.4 Algemene voorwaarden voor calibratie in onvolledige designs

In deze paragraaf zullen we ingaan op de algemene voorwaarden die moeten gelden voor het bestaan van eindige en unieke itemparameterschattingen voor zowel de CML- als de MML-methode in onvolledige designs. We bespreken hier in feite alleen de voorwaarden die moeten gelden in gefixeerde onvolledige designs, waarbij de onderzoeker het ontstaan van de onvolledige gegevens volledig onder controle heeft. Zie de itemindicator verdeling (6.3). In paragraaf 6.5 gaan we dan in op de nadere voorwaarden die gesteld moeten worden aan een calibratiemethode bij stochastische designs.

In gefixeerde onvolledige designs geldt voor de calibratie, met welke methode dan ook, dat het in ieder geval noodzakelijk is dat er tussen de verschillende te onderscheiden volledige deeldesigns iets gemeenschappelijk is. In paragraaf 6.1 werd al aangegeven dat dit nodig is om in een onvolledig design de itemparameters op één schaal te kunnen brengen. Om ervan verzekerd te zijn voor alle parameters unieke schattingen te krijgen moet deze voorwaarde nog iets worden aangescherpt. In de psychometrische literatuur zijn de voorwaarden voor het bestaan van en het uniek zijn van CML-schattingen in gefixeerde onvolledige designs in het Raschmodel exact uitgewerkt door Fischer (1981). Omdat de voorwaarden aan het design voor het bestaan van CML-schattingen strenger zijn dan voor het bestaan van MML-schattingen, zullen we deze hierna kort schetsen. Over de minder strenge condities aan het design bij MML zullen we daarna enkele opmerkingen maken.

Fischer (1981) toont in eerste instantie aan onder welke voorwaarden er eindige en unieke CML-schattingen voor de itemparameters in volledige designs bestaan, waarna hij zijn resultaten generaliseert naar het bestaan en uniek zijn van de schattingen in onvolledige designs. We geven nu, zonder op details in te gaan, een beschrijving van deze voorwaarden. In volledige designs worden Fischers voorwaarden gesteld aan de datamatrix van alle itemantwoorden:

$$\boldsymbol{x} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & \cdots & x_{nk} \end{bmatrix}.$$

De rij-index van deze matrix geeft een persoon aan, de kolom-index een item. Om itemparameterschattingen te verkrijgen is het noodzakelijk dat de kolomsommen uit

deze matrix $t_j = \sum_{v=1}^{n} x_{vj}$ niet gelijk zijn aan 0, iedereen maakt de opgave fout, of aan n, iedereen maakt de opgave goed. Zoals we in hoofdstuk 4 zagen bereikt de aannemelijkheidsfunctie voor zo'n item zijn maximum bij respectievelijk $-\infty$ en ∞ en bestaat er dus geen eindige schatting van de itemparameter voor dat item. Aan deze voorwaarde moet voor elk item $j = 1,\ldots,k$ voldaan zijn. Fischer geeft aan dat voor de gehele datamatrix $\boldsymbol{x}$ nog iets meer moet gelden: het mag niet zo zijn dat deze uiteenvalt in twee delen die geen verbinding met elkaar hebben. Hij definieert daarvoor het begrip 'goed geconditioneerd' zijn van de datamatrix en toont aan dat het goed geconditioneerd zijn van de datamatrix de voorwaarde is voor het bestaan van unieke schattingen van de itemparameters. Een datamatrix is goed geconditioneerd als in elke mogelijke opdeling van de items in twee niet-lege deelverzamelingen I_1 en I_2 er minstens één persoon is die een item uit I_1 goed heeft en een item uit I_2 fout heeft. Anders heet de datamatrix 'slecht geconditioneerd'.

Stel we hebben een opdeling van de items, I_1 en I_2. Dan kunnen we de personen proberen op te delen in drie groepen: P_1 bestaat uit de personen die alle items uit deelverzameling I_2 goed hebben; P_2 bestaat uit alle personen die alle items uit deelverzameling I_1 fout hebben met uitzondering van de personen die al in groep P_1 zitten; de groep personen P_3 zijn alle personen die niet in groep P_1 of P_2 zitten. Dan kunnen we door permutaties van rijen en kolommen de datamatrix altijd schrijven als

$$\boldsymbol{x} = \begin{bmatrix} [x^1] & [x^2] \\ [x^3] & [x^4] \\ [x^5] & [x^6] \end{bmatrix} = \begin{array}{c} \\ \\ P_1 \\ \\ P_2 \\ \\ P_3 \end{array} \begin{array}{c} I_1 \qquad\quad I_2 \\ \begin{bmatrix} & & 1 & \ldots & 1 \\ & [x^1] & \ldots & \ldots & \ldots \\ & & 1 & \ldots & 1 \\ 0 & \ldots & 0 & & \\ \ldots & \ldots & \ldots & [x^4] & \\ 0 & \ldots & 0 & & \\ & [x^5] & & [x^6] & \end{bmatrix} \end{array} .$$

Hierin staan de zes submatrices $[x^l], l = 1,\ldots,6$, de niet gespecificeerde matrices bevatten in principe rijen en kolommen waarin niet alleen 0 of alleen 1 staat. Fischer toont aan dat als er voor een datamatrix een opdeling van de items bestaat waarvoor de submatrices $[x^5]$ en $[x^6]$ leeg zijn, ofwel dat er voor de datamatrix voor die

opdeling van de items geen enkele persoon in groep P_3 zit, dan is de datamatrix slecht geconditioneerd. De datamatrix is goed geconditioneerd als er voor elke opdeling in de items I_1 en I_2 er op zijn minst een persoon in groep P_3 zit. Dat willen zeggen in $\left[x^5\right]$ zit minstens een rij met niet alleen 0 en in $\left[x^6\right]$ en rij met niet alleen 1. Het formele bewijs van Fischer zullen we hier niet reproduceren. Echter het idee achter het goed geconditioneerd moeten zijn van de matrix voor de schatting van de parameters en dus dat het voor een datamatrix noodzakelijk is dat de derde groep P_3 bestaat, is als volgt. Zou de derde groep niet bestaan dan kan aangetoond worden dat de aannemelijkheidsfunctie blijft stijgen als de parameterwaarden van de items in I_2 steeds groter worden; voor de items in I_1 is dat het geval als de parameterwaarden steeds kleiner worden. Er bestaan dan met andere woorden geen eindige schatters. Het bestaan van P_3 brengt de noodzakelijke verbinding in de datamatrix tot stand die dit voorkomt.

De voorwaarden voor het eindig en uniek zijn van CML-schattingen in onvolledige designs in het Raschmodel zijn hetzelfde (Fischer, 1981) met dien verstande dat de submatrices $[x^2]$ en $\left[x^3\right]$ behalve respectievelijk enen en nullen ook lege cellen mag bevatten. De lege cellen duiden dan de ontbrekende itemantwoorden aan. Op analoge wijze kan dan goed geconditioneerd zijn van de datamatrix gedefinieerd worden en kan worden aangetoond dat dit ook de voorwaarde voor het eindig en uniek zijn van de schattingen is. Fischer (1981) geeft een eenvoudige algoritme om de vervulling van deze conditie na te gaan. Tenslotte zij nog opgemerkt dat in de praktijk doorgaans aan de voorwaarden is voldaan als een anker bestaat uit een tiental niet te extreme opgaven.

Als bij het Raschmodel aan de CML-voorwaarden aan de datamatrix is voldaan dan leert de praktijk dat dan tevens aan de voorwaarden voor het bestaan van de parameterschattingen bij MML is voldaan. Hierbij moeten we echter wel bedenken dat bij CML (zie hoofdstuk 4) geen enkele aanname behoeft te worden gedaan omtrent de vaardigheid van de steekproeven leerlingen waarmee we items calibreren. Bij MML echter hebben we expliciet de aanname nodig dat de steekproef waarmee we onze items calibreren, een aselecte is uit één en dezelfde gespecificeerde verdeling, waarvan we de parameters gelijk met de itemparameters schatten. Dan wel dat we aselecte steekproeven hebben uit meerdere verdelingen, waarbij we van elke verdeling parameters schatten samen met de itemparameters (zie paragraaf 4.4). Als aan deze extra aanname is voldaan dan behoeft de verbondenheidsvoorwaarde bij MML niet meer te gelden. De verbinding kan dan worden gevonden in de equivalente groepen personen die verschillende items maken.

Over de toepasbaarheid van de CML- en de MML-schattingsmethode in de bij onvolledige designs behorende datamatrices, zoals die in paragraaf 6.2 besproken zijn,

kunnen we op basis van het bovenstaande in het algemeen het volgende concluderen. De datamatrices van het niet-verbonden design en het gedeeltelijk verbonden design kunnen niet gecalibreerd worden met de CML-methode en eventueel (met de extra aanname) wel met de MML-methode. De overige matrices komen in principe voor beide in aanmerking.

Tenslotte zij opgemerkt dat de bestaansvoorwaarden voor CML- en MML-schattingen in onvolledige designs, zoals hiervoor besproken slechts betrekking hebben op het Raschmodel. Voor uitgebreidere modellen, zoals het OPLM en voor modellen met polytome items, zijn de voorwaarden uiteraard complexer. Generaliseringen van het voorgaande voor deze modellen zijn mogelijk, maar deze zullen we niet bespreken.

6.5 Voorwaarden voor calibratie in stochastische designs

In deze paragraaf gaan we ervan uit dat aan de algemene voorwaarden uit paragraaf 6.4 is voldaan en zullen we beschrijven aan welke extra voorwaarden moet worden voldaan voor calibratie van de items in gerandomiseerde, in meerfasen en in groepsgerichte designs. We zullen daarbij opnieuw onderscheid maken tussen CML en MML als calibratie methode. In onze voorbeelden beperken we ons hierbij opnieuw tot het Raschmodel, echter de principes die besproken worden, kunnen ook op de in hoofdstuk 5 besproken uitgebreidere modellen worden toegepast.

De eerste centrale vraag die we bij alle stochastische designs moeten beantwoorden is: moeten we bij de analyse van de gegevens altijd rekening met het stochastische karakter van de designvariabele zelf of kunnen we in de analyse de designvariabele evengoed negeren, zonder dat dit gevolgen heeft voor de analyse. Voor de goede orde zij opgemerkt, dat we met het negeren van de designvariabele in de analyse bedoelen dat het stochastische karakter ervan in de analyse buiten beschouwing wordt gelaten; de informatie wie welke items heeft gemaakt kan natuurlijk nooit worden genegeerd. Het is voor te stellen dat de mogelijkheid om de designvariabele buiten de analyse te houden de analyse soms veel simpeler kan maken. Als we rekening moeten houden met de toevalsstructuur van het design, dan hebben we niet alleen de itemantwoorden X_{vi} als toevalsvariabelen, maar ook het al of niet hebben van dat antwoord. Of anders geformuleerd, als we bijvoorbeeld een aannemelijkheidsfunctie beschouwen dan kijken we bij het negeren van de designvariabele slechts naar de verdeling van alle geobserveerde itemantwoorden $\boldsymbol{X}_{obs}$, terwijl we bij het meenemen van de design-variabele de simultane verdeling van $(\boldsymbol{X}_{obs}, \boldsymbol{R})$ zullen moeten beschouwen. Door Rubin (1976) is een algemene theorie ontwikkeld met betrekking tot de analyse met

ontbrekende gegevens, waarin het eventueel negeren van de designvariabele centraal staat. Zijn begrippenkader, dat later met meer voorbeelden is uitgewerkt in Little en Rubin (1987), is in de itemresponstheorie onder meer door Mislevy en Whu (1988), Mislevy en Sheenan (1989) en door Eggen en Verhelst (1992) gehanteerd om analyse mogelijkheden in stochastische designs te beschrijven. De laatsten geven zowel voor de CML- als de MML-methode de voorwaarden voor calibratie in de drie genoemde designs.

In het hiernavolgende zullen we voornamelijk de resultaten van Eggen en Verhelst (1992) samenvatten en met voorbeelden illustreren. Alvorens dit te doen zullen we echter twee onderwerpen nog nader moeten bespreken. Het betreft allereerst het begrippenkader van Rubin (1976) en vervolgens de voor de calibratie in onvolledige designs essentiële verschillen tussen de CML- en de MML-schattingsmethode. Eerst echter een opmerking over het grote praktische belang van de mogelijkheid het design te negeren in de IRT. Belangrijk is dat in de IRT de standaardprogrammatuur die ontwikkeld is voor zowel de CML- als MML-analyse impliciet uitgaat van het negeren van de designvariabele in de analyse. Data afkomstig uit niet-negeerbare designs kunnen dus niet geanalyseerd worden met standaardprogrammatuur. In de praktijk is het echter zo dat aan de data niet te 'zien' is uit welk design ze komen. Dat wil zeggen, de programmatuur behandelt ze alsof ze uit negeerbare designs komen en levert in het geval het design niet negeerbaar is onjuiste uitkomsten. Het belang van het voldaan zijn aan de voorwaarden voor het negeren van het stochastische karakter van het design is daarom evident om foute resultaten te voorkomen.

Rubins theorie

Rubin introduceert het zogenaamde 'ignorability' principe. Dit principe wordt onder andere gedefinieerd voor statistische analyse met de grootste-aannemelijkheid ofwel ML-methode (Maximum Likelihood). Omdat de calibratie van items, en trouwens ook het schatten van persoonsparameters, in IRT plaatsvindt met deze methode zullen we de voorwaarden voor correct toepassen van dit principe hiertoe beperken. Dit principe houdt in dat we ons voor de analyse van gegevens kunnen beperken tot slechts de resultaten op waargenomen variabelen, zonder dat we in de procedure ook informatie over het design moeten meenemen. Het design wordt genegeerd. In het algemeen beschouwen we in een analyse een vector toevalsvariabele $\boldsymbol{U} = (U_1,...,U_m)$ met verdeling $f_\tau(\boldsymbol{u})$. De parametervector τ bevat de parameters die we willen schatten. Om de gedachten te bepalen is het voor te stellen dat $m = n.k$, met k het aantal

variabelen en n het aantal personen dat in de analyse wordt beschouwd. Als er ontbrekende gegevens zijn, definiëren we een 'missing data indicator' $\boldsymbol{M} = (M_1,...,M_m)$, die aangeeft of een variabele U_j daadwerkelijk geobserveerd is, $m_j = 1$, of niet, $m_j = 0$. Dus $\boldsymbol{M}$ is op dezelfde wijze gedefinieerd als de itemindicator variabele $\boldsymbol{R}$ in paragraaf 6.3. $\boldsymbol{M}$ wordt echter, zoals verderop duidelijk zal worden algemener gebruikt dan alleen als itemindicator $\boldsymbol{R}$. De missing data indicator partitioneert $\boldsymbol{U}$ en zijn geobserveerde waarde $\boldsymbol{u}$ in

$$\boldsymbol{U} = (\boldsymbol{U}_{obs}, \boldsymbol{U}_{mis}) \text{ en } \boldsymbol{u} = (\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}). \tag{6.6}$$

De verzameling *obs* bevat de indexen van waargenomen variabelen, dat wil zeggen, elke j waarvoor $m_j = 1$, en *mis* is de verzameling van indexen van de niet waargenomen variabelen $(m_j = 0)$. $\boldsymbol{U}_{obs}$ en $\boldsymbol{u}_{obs}$ zijn respectievelijk de toevalsvariabele en de realisatie van de waargenomen variabelen. $\boldsymbol{U}_{mis}$ de toevalsvariabele en $\boldsymbol{u}_{mis}$ de waarden die we geobserveerd zouden hebben, als we dat gewild of gekund hadden, van de niet waargenomen variabelen. In een analyse met de grootste-aannemelijkheidsmethode zouden we ons moeten baseren op de gezamenlijke verdeling $g_{\tau,\phi}$ van alle waargenomen toevalsvariabelen, dat wil zeggen van $\boldsymbol{U}_{obs}$ en $\boldsymbol{M}$:

$$g_{\tau,\phi}(\boldsymbol{u}_{obs}, \boldsymbol{m}) = \int_{\boldsymbol{u}_{mis}} g_{\tau,\phi}(\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}, \boldsymbol{m}) \quad d\boldsymbol{u}_{mis}. \tag{6.7}$$

We merken op dat we in het hoofdstuk een uitdrukking als (6.7) zowel voor een verdeling van toevalsvariabele gebruiken als voor een aannemelijkheidsfunctie, zonder dat laatste expliciet als functie van de parameter(s) te schrijven. In (6.7) staat ϕ voor een mogelijke parameter van de verdeling van de missing data indicator $\boldsymbol{M}$. Bij n personen en experimentele onafhankelijkheid (zie hoofdstuk 4) is dit ook te schrijven als:

$$\int_{\boldsymbol{u}_{mis}} g_{\tau,\phi}(\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}, \boldsymbol{m})\, d\boldsymbol{u}_{mis} = \prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} g_{\tau,\phi}(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v}, \boldsymbol{m}_v)\, d\boldsymbol{u}_{mis,v}. \tag{6.8}$$

We zien dat (6.8) zowel afhangt van de verdeling van $\boldsymbol{M}$, met parameter ϕ, als van de variabele $\boldsymbol{U}$, met parameter τ, waarin we geïnteresseerd zijn. Als we in plaats van (6.8)

$$\int_{\boldsymbol{u}_{mis}} f_\tau(\boldsymbol{u})\, d\boldsymbol{u}_{mis} \;=$$

$$\int_{\boldsymbol{u}_{mis}} f_\tau(\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis})\, d\boldsymbol{u}_{mis} \;= \tag{6.9}$$

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v}\,,$$

zouden toepassen, dan negeren we de designvariabele in de analyse. We hebben dan een eenvoudiger uitdrukking die alleen afhangt van de verdeling van de variabelen die ons interesseren, met parameter τ. Als het geoorloofd is, dat wil zeggen niet tot fouten leidt, (6.9) in plaats van (6.8) in de analyse toe te passen dan geldt het 'ignorability' principe. Zonder fouten te maken nemen we dan aan dat de observaties van $\boldsymbol{U}$ uit de marginale verdeling van alleen de waargenomen variabelen $\boldsymbol{U}_{obs}$ komen en we negeren de designvariabele. De rechtvaardiging hiervan hangt af van de eigenschappen die de verdeling van de missing data indicator heeft, of zoals Rubin het noemt: van de eigenschappen van "the proces that causes missing data". Dit proces wordt door Rubin beschreven met de voorwaardelijke verdeling van de missing data indicator gegeven de data: $h_\phi(\boldsymbol{m} \mid \boldsymbol{u})$. Als voor deze verdeling de eigenschap geldt dat

$$h_\phi(\boldsymbol{m} \mid \boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}) = h_\phi(\boldsymbol{m} \mid \boldsymbol{u}_{obs}) \text{ voor alle } \boldsymbol{u}_{mis}, \tag{6.10}$$

dan is het gerechtvaardigd het design in de ML-analyse te negeren. Ofwel de kansen op het ontbreken van de gegevens hangen niet af van de waarden van de gegevens die niet zijn waargenomen, maar hangen mogelijkerwijs uitsluitend af van wel waargenomen gegevens. Rubin noemt de situatie waarin dit geldt MAR, 'missing at random'. We tonen nu aan dat als aan de MAR-voorwaarde (6.10) voldaan is, we in de ML-analyse evengoed uit kunnen gaan van de eenvoudiger verdeling (6.9) als van (6.8). Het rechterlid van (6.8) kunnen we in het algemeen herschrijven als:

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} g_{\tau,\phi}(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v}, \boldsymbol{m}_v)\, d\boldsymbol{u}_{mis,v} =$$

(6.11)

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v}) \cdot f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v} =$$

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v}) \cdot f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v} =$$

$$\prod_{v=1}^{n} h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v}) \cdot \prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v}$$

In (6.11) maken we in de eerste gelijkheid gebruik van de eigenschappen van voorwaardelijke kansen: de gezamenlijke verdeling $g_{\tau,\phi}$ wordt geschreven als het produkt van de voor-waardelijke verdeling h_ϕ van de missing data indicator en de verdeling van dat deel waarop geconditioneerd wordt. Deze laatste verdeling is de verdeling f_τ van de variabelen $\boldsymbol{u} = (\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis})$. In de volgende gelijkheid wordt gebruik gemaakt van de MAR-eigenschap (6.10) van de verdeling van de designvariabele. Omdat tenslotte $h_\phi(\boldsymbol{m}_v \mid \boldsymbol{u}_{obs,v})$ onafhankelijk is van $\boldsymbol{u}_{mis,v}$ kan deze term buiten de integraal worden gehaald. Het resultaat is dat de aannemelijkheidsfunctie (6.8) uiteenvalt in twee termen, waarvan de tweede term gelijk is aan de eenvoudigere aannemelijkheidsfunctie (6.9) en een eerste term die onafhankelijk is van de parameter τ waarnaar we de aannemelijkheidsfunctie moeten maxi-maliseren. Het zal duidelijk zijn dat we bij het maximaliseren naar τ deze eerste term evengoed kunnen weglaten. Voor de goede orde zij vermeld dat naast de MAR-voorwaarde ook nog voldaan moet zijn aan een voorwaarde, die betrekking heeft op de mogelijke waarden die de te schatten parameters τ en eventuele parameters ϕ van de verdeling van de missing data indicator kunnen aannemen. Aangezien aan deze voorwaarde in onze toepassing altijd voldaan is, zullen we hieraan geen aandacht besteden. Aldus hebben we gezien dat het voldoen aan de MAR-voorwaarde voldoende is voor het negeren van het design in de analyse.

Soms geldt dat de ontbrekende gegevens MCAR, 'missing completely at random', zijn, hetgeen betekent dat

$$h_\phi(\boldsymbol{m} \mid \boldsymbol{u}_{obs}, \boldsymbol{u}_{mis}) = h_\phi(\boldsymbol{m}) \text{ voor alle } \boldsymbol{u}_{mis} \text{ en } \boldsymbol{u}_{mis}. \tag{6.12}$$

Dat wil zeggen de kans op het ontbreken van gegevens hangt noch van de waargenomen noch de niet waargenomen gegevens af. Het zal duidelijk zijn dat als aan de sterkere MCAR voorwaarde is voldaan automatisch ook voldaan aan de MAR-voorwaarde.

Verschil designvariabele bij CML en MML

In de analyse in onvolledige designs verschillen de CML- en MML-schattingsmethode op een essentieel punt van elkaar. De reden voor het onderscheid tussen CML en MML is dat in genoemde designs het mechanisme dat verantwoordelijk voor het ontbreken van gegevens een toevalsproces is en dat bij de calibratie met de CML- en MML-methode er in principe uitgegaan wordt van een verschillend toevalsproces dat de itemantwoorden genereert. Bij CML worden alleen de itemantwoorden $X_{vi}, v = 1,...,n; i = 1,...,k$ als toevalsvariabelen beschouwd, terwijl bij MML naast deze itemantwoorden ook de vaardigheden van de personen die items maken $\theta_v, v = 1,...,n$ expliciet als toevalsvariabelen worden beschouwd. De consequentie hiervan is dat de algemene missing data indicator voor een persoon $\boldsymbol{M}_v$ bij MML altijd één element meer bevat dan bij CML. Als de totale itemverzameling bijvoorbeeld vijf items bevat, waarvan een bepaald persoon v, volgens een of ander stochastisch design uit paragraaf 6.3, het eerste, het derde en het vierde item wel maakt en de andere twee items niet dan heeft de missing data indicator bij een CML-analyse dezelfde waarde als de itemindicator $\boldsymbol{m}_v = \boldsymbol{r}_v = (1,0,1,1,0)$. In de MML-analyse daarentegen is $\boldsymbol{m}_v = (\boldsymbol{r}_v, 0) = (1,0,1,1,0,0)$, waarin de laatste 0 het niet waarnemen van de variabele θ_v indiceert.

In Eggen en Verhelst (1992) is uiteengezet, dat Rubins voorwaarden voor het negeren van de designvariabelen in de analyse bij de MML-methode onverkort toepasbaar zijn. Het controleren van Rubins voorwaarden geeft uitsluitsel over de mogelijkheid de designvariabele te negeren in de analyse. In paragraaf 6.5.1, zullen we dit voor de stochastische designs uit paragraaf 6.3 bespreken. Bij CML blijken Rubins voorwaarden niet beslissend te zijn. De mogelijkheid van toepassing van CML in stochastische designs blijkt in de eerste plaats af te hangen van dat deel van de aannemelijkheidsfunctie dat we in de CML-analyse buiten beschouwing laten. In paragraaf 6.5.2 zullen we dat uitwerken. In deze paragrafen zullen wij als in hoofdstuk 4 een deel van de uitwerkingen alleen geven voor het Raschmodel, de principes zijn echter evenzeer toepasbaar voor de uitgebreidere modellen die in hoofdstuk 5 zijn behandeld. De verdeling van de itemantwoorden, ook als we deze als aannemelijkheids-functie beschouwen, zullen we daarbij steeds aangeven met $P_{\cdot\cdot}(...)$.

6.5.1 MML in stochastische designs

Aansluitend bij de notatie in hoofdstuk 4 en uit de vorige paragraaf hebben we in een MML-analyse te maken met de toevalsvariabele

$$\boldsymbol{U} = (\boldsymbol{X}, \theta) = (\boldsymbol{X}_1, \theta_1, ..., \boldsymbol{X}_n, \theta_n). \tag{6.13}$$

Met θ_v de vaardigheid van persoon v, $v = 1,...,n$ en $\boldsymbol{X}_v = (X_{v1},..., X_{vk})$ de antwoorden van deze personen op de k items, die eventueel niet allemaal zijn geobserveerd. De parametervector die we willen schatten is $\tau = (\beta, \mu, \sigma^2)$, met $\beta = (\beta_1,..., \beta_k)$ de vector van alle k moeilijkheidsparameters en μ en σ^2, respectievelijk het gemiddelde en de variantie van de normale vaardigheidsverdeling $g_{\mu,\sigma^2}(\theta)$ (zie formule 4.55).

MML in gerandomiseerde onvolledige designs.

In deze designs is de verdeling van de missing data indicator gelijk aan de verdeling van de itemindicator (zie (6.2)), omdat de vaardigheid θ_v nooit wordt waargenomen geldt:

$$P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0)) = P(\boldsymbol{R}_v = \boldsymbol{r}_b) = \phi_b. \tag{6.14}$$

Hierin is $\boldsymbol{r}_b$ zoals eerder de vector met lengte k met een 1 op de plaatsen die de items indiceren die in boekje b zitten en een 0 op de overige plaatsen. Deze formule geldt uiteraard weer voor alle personen: $v = 1,...,n$ en alle boekjes $b = 1,...,B$.

Als we kijken waarin de totale verzameling van toevalsvariabelen $\boldsymbol{U}$ (6.13) uiteenvalt door de missing data indicator $\boldsymbol{M}_v$ volgens (6.6), dan is eenvoudig na te gaan dat in dit geval voor elke persoon v geldt:

$$\left.\begin{aligned} \boldsymbol{U}_{obs,v} &= \boldsymbol{X}_{obs,v} \\ \boldsymbol{U}_{mis,v} &= (\boldsymbol{X}_{mis,v}, \theta_v) \end{aligned}\right\}, \quad (v = 1,...,n). \tag{6.15}$$

In (6.14) zien we dat de verdeling van de missing data indicator noch van de waarden van de niet waargenomen data noch van de waargenomen data afhangt. De ontbrekende data zijn in gerandomiseerde designs dus MCAR, formule (6.12) is geldig, en duidelijk is dat aan Rubins voorwaarden voor het negeren van het design is voldaan. Het bewijs hiervan, een toepassing van (6.11) laten we aan de lezer over. We kunnen dus de marginale verdeling van de observaties $\boldsymbol{X}_{obs}$ als basis voor de analyse

gebruiken. De aannemelijkheidsfunctie wordt dan gegeven door het in (6.9) invullen van de specificatie (6.15):

$$\prod_{v} \int_{x_{mis,v}} \int_{\theta_v} f_\tau(\boldsymbol{x}_{obs,v}, \boldsymbol{x}_{mis,v}, \theta_v)\, d\theta_v\, d\boldsymbol{x}_{mis,v} =$$

$$\prod_{v} \int_{x_{mis,v}} \int_{\theta_v} P_\beta(\boldsymbol{x}_{obs,v}, \boldsymbol{x}_{mis,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v\, d\boldsymbol{x}_{mis,v} = \tag{6.16}$$

$$\prod_{v} \int_{\theta_v} P_\beta(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v \ .$$

In (6.16) volgt de eerste gelijkheid uit de eigenschappen van voorwaardelijke kansen, zoals we die eerder bij de afleiding van de marginale aannemelijkheidsfunctie, formule (4.49), zagen. De tweede gelijkheid volgt uit de lokale stochastische onafhankelijkheid van de itemantwoorden en het uitintegreren van $\boldsymbol{x}_{mis,v}, v = 1,...,n$. De aannemelijkheidsfunctie (6.16) lijkt uiteindelijke dus zeer veel op de marginale aannemelijkheidsfunctie voor volledige gegevens (formule 4.57). Het verschil zit er slechts in dat per persoon v slechts de kansen op de waargenomen responsen worden meegenomen en dat per persoon alleen de itemparameters van de waargenomen items in de aannemelijkheidsfunctie meedoen. De relatie met de volledige data MML-analyse wordt duidelijk gemaakt als we met n_b het aantal personen noteren dat boekje b maakt, dan geldt dat $\sum_{b=1}^{B} n_b = n$, het totaal aantal personen. Als we verder $\beta_{(b)}$ definiëren als de k_b-vector van de itemparameters van de items in boekje b, dan kunnen we (6.16) herschrijven als

$$\prod_{v=1}^{n} \int_{\theta_v} P_\beta(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v =$$

$$\prod_{b=1}^{B} \prod_{v=1}^{n_b} \int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu,\sigma^2}(\theta_v)\, d\theta_v . \tag{6.17}$$

We zien in (6.17) dus dat we de marginale aannemelijkheidsfunctie in onvolledige designs kunnen schrijven als een produkt van B marginale aannemelijkheidsfuncties, evenveel als er verschillende toetsboekjes zijn, voor volledige gegevens. Vergelijk formule (4.113).

MML in meerfasen onvolledige designs

In meerfasen designs is de opdeling door de missing data indicator in geobserveerde en niet geobserveerde variabelen hetzelfde als bij gerandomiseerde designs (zie (6.15)). De verdeling van de missing data indicator volgt op dezelfde wijze als bij gerandomiseerde designs nu echter met de itemindicator van meerfasen designs (6.4) als basis:

$$P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0) \mid \boldsymbol{x}_{obs,v}) = P(\boldsymbol{R}_v = \boldsymbol{r}_b \mid \boldsymbol{x}_{obs,v}) = 0 \quad \text{of} \quad 1. \tag{6.18}$$

Formule (6.18) geldt voor elke persoon $v = 1,...,n$ en elk boekje $b = 1,...,B$. Eenvoudig is in te zien dat de verdeling van de missing data indicator voldoet aan de voorwaarde (6.10), dat wil zeggen de missing data zijn MAR. De designverdeling hangt immers alleen af van de geobserveerde waarden en niet van de niet geobserveerde. Volgens het ignorability principe is het dus gerechtvaardigd het design in de analyse te negeren. De algemene uitdrukking voor de marginale aannemelijkheidsfunctie is in dit geval identiek aan de marginale aannemelijk-heidsfunctie bij gerandomiseerde designs (6.16) of (6.17).

In paragraaf 6.5.2 zullen we in tabel 6.6. een voorbeeld van een MML-analyse in een meerfasen design geven en de resultaten vergelijken met een CML-analyse.

MML in groepsgerichte designs

In groepsgerichte calibratiedesigns hebben we in paragraaf 6.3.3 twee situaties onderscheiden. In de eerste hebben wij een achtergrondvariabele Y die slechts een rol speelt in de toewijzing van boekjes aan leerlingen en zijn we niet geïnteresseerd in de verschillende vaardigheids-verdelingen. In de tweede zijn we behalve in de itemparameters ook geïnteresseerd in de parameters van de in totaal B vaardigheidsverdelingen voor de verschillende niveaus van de achtergrondvariabele: we kunnen B subpopulaties onderscheiden in de totale populatie. In de tweede situatie zullen we in de praktijk vaak niet één aselecte steekproef uit een vaardigheids-verdeling ter beschikking hebben, maar, een bewust op die wijze getrokken gestratificeerde steekproef, bestaande uit aselecte steekproeven uit de vaardigheidsverdelingen voor elk onderscheiden niveau van de achtergrondvariabele.

Hetzelfde mogelijke onderscheid in subpopulaties speelt ook al een rol bij de MML-analyse in volledige designs. Bij een gestratificeerde steekproef zullen we daar, samen

met de itemparameters, de parameters van meer vaardigheidsverdelingen moeten schatten. Als we dat niet zouden doen, en de steekproef beschouwen als een aselecte uit één populatie, dan maken we een specificatiefout welke tot onjuiste schattingen leidt. Aangezien de situatie van volledige designs een bijzonder geval van is groepsgerichte designs, zullen we hieraan verder geen expliciet aandacht besteden.

Mislevy en Sheenan (1989) hebben aangetoond dat het voor de behandeling van de designvariabele in groepsgerichte designs in een MML-analyse niet uitmaakt of we nu een aselecte steekproef hebben uit één populatie of een gestratificeerde. Vandaar dat we er in deze paragraaf van uit zullen gaan dat we een aselecte steekproef hebben uit één vaardigheids-verdeling, die kan worden geschreven als een combinatie van B verdelingen, voor elke subpopulatie geassocieerd met een onderscheiden niveau van de achtergrondvariabele Y:

$$\begin{aligned} g_{\mu,\sigma^2}(\theta) &= \sum_{b=1}^{B} P(\theta, Y = y_b) = \sum_{b=1}^{B} P(\theta \mid Y = y_b).\ P(Y = y_b) \\ &= \sum_{b=1}^{B} g_{\mu_b,\sigma_b^2}(\theta).\pi_b \quad . \end{aligned} \tag{6.19}$$

In (6.19) zijn μ_b en σ_b^2 het gemiddelde en de variantie van de vaardigheidsverdeling verdeling in subpopulatie b en π_b de proportie personen in subpopulatie b in de totale populatie.

In groepsgerichte designs is de verdeling van de itemindicator gegeven in (6.5), waaruit met (6.19) volgt dat

$$P(\boldsymbol{R}_v = \boldsymbol{r}_b) = P(Y_v = y_b) = \pi_b.$$

Hetgeen uiteraard weer geldt voor alle personen $v = 1,...,n$ en alle boekjes of onderscheiden niveaus $b = 1,...,B$ van de achtergrondvariabele. Omdat de vaardigheid θ_v nooit geobserveerd wordt komt de vraag of we in deze designs de designvariabele kunnen negeren neer op de vraag of we in de analyse de achtergrondvariabele Y kunnen negeren ofwel moeten meenemen. Het antwoord op deze vraag kunnen we weer geven door de voorwaarden van Rubin te controleren.

In de MML-analyse zijn in dit geval de toevalsvariabelen die een rol zouden kunnen spelen $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{Y}, \theta)$, met voor elke persoon de vector $\boldsymbol{X}_v$ met antwoorden op de k items, de waarde van de achtergrondvariabele Y_v en de vaardigheid θ_v. Als we de

achtergrond-informatie in de analyse meenemen dan wordt de opdeling van $\boldsymbol{U}$ door de missing data indicator $\boldsymbol{M}_v$ gegeven door

$$\left.\begin{aligned} \boldsymbol{U}_{obs,v} &= (\boldsymbol{X}_{obs,v}, Y_v) \\ \boldsymbol{U}_{mis,v} &= (\boldsymbol{X}_{mis,v}, \theta_v) \end{aligned}\right\}, \quad (v = 1,...,n). \tag{6.20}$$

En de verdeling van $\boldsymbol{M}_v$ door

$$P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1, 0)) = P(\boldsymbol{R}_v = \boldsymbol{r}_b) = P(Y_v = y_b),$$

ofwel

$$\left.\begin{aligned} P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1, 0) \mid Y_v = y_b) &= 1 \\ P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1, 0) \mid Y_v \neq y_b) &= 0 \end{aligned}\right\}, \quad b = 1,..., B; v = 1,...,n. \tag{6.21}$$

Waarbij de waarde 1 van het voorlaatste element van $\boldsymbol{M}_v$ aanduidt dat Y_v als waargenomen wordt beschouwd en het laatste element het niet waarnemen van θ_v indiceert. Uit (6.21) is eenvoudig te zien dat bij het meenemen van de achtergrondvariabele aan de MAR-voorwaarde (6.10) is voldaan: de verdeling van de missing data indicator hangt alleen af van geobserveerde waarden, en in de analyse kunnen we de designvariabele als geheel negeren en de marginale verdeling van alleen de geobserveerde waarden (6.9) hoeven we te beschouwen. Als we de kans beschouwen dat een aselect getrokken persoon uit de populatie een bepaald antwoordpatroon heeft in boekje b, dan kunnen we met de eerdere notatie (formule (6.17)) hiervoor schrijven:

$$\begin{aligned} &P_{\beta_{(b)}, \mu_b, \sigma_b^2, \pi_b}(\boldsymbol{x}_{obs,v}, Y_v = y_b) = \\ &\int_{\boldsymbol{x}_{mis,v}} \int_{\theta_v} P_{\beta_{(b)}, \mu_b, \sigma_b^2, \pi_b}(\boldsymbol{x}_{obs,v}, \boldsymbol{x}_{mis,v}, Y_v = y_b, \theta_v)\ d\theta_v\ d\boldsymbol{x}_{mis,v} = \\ &\int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid Y_v = y_b, \theta_v).\ P_{\mu_b, \sigma_b^2}(\theta_v \mid Y_v = y_b).P_{\pi_b}(Y_v = y_b)\ d\theta_v = \\ &\pi_b . \int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid \theta_v).g_{\mu_b, \sigma_b^2}(\sigma_v)\ d\theta_v . \end{aligned} \tag{6.22}$$

De tweede gelijkheid in (6.22) volgt uit de eigenschappen van voorwaardelijke kansen, terwijl in de derde gebruik gemaakt wordt van de lokale stochastische onafhankelijkheid in IRT-modellen. Bij n_b personen die boekje b maken wordt de marginale aannemelijkheidsfunctie gegeven door:

$$\prod_{b=1}^{B} \pi_b^{n_b} . \prod_{b=1}^{B} \prod_{v=1}^{n_b} \int_{\theta_v} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid \theta_v) . g_{\mu_b, \sigma_b^2}(\theta_v)\, d\,\theta_v . \tag{6.23}$$

We zien dat (6.23) uiteenvalt in een deel dat alleen afhangt van de trekkingskansen π_b, dat een persoon uit subpopulatie b komt en een deel dat het produkt is van in totaal B deels overlappende marginale aannemelijkheidsfuncties als (4.57). Voor de schatting van de parameters kunnen we deze functie maximaliseren naar β, μ_b, σ_b^2 en eventueel π_b, voor $b = 1,...,B$. De ML-schatter van π_b is gegeven door: $\hat{\pi}_b = n_b/n$.

Als we in groepsgerichte designs de achtergrondvariabele Y_i niet zouden meenemen dan wordt de opdeling van $\boldsymbol{U}$ gegeven door (vergelijk met (6.20))

$$\left.\begin{aligned} \boldsymbol{U}_{obs,v} &= \boldsymbol{X}_{obs,v} \\ \boldsymbol{U}_{mis,v} &= (\boldsymbol{X}_{mis,v}, Y_v, \theta_v) \end{aligned}\right\}, \quad (v = 1,...,n).$$

Immers Y_v beschouwen we dan als niet waargenomen gegevens. De verdeling van $\boldsymbol{M}_v$ is dan (vergelijk met (6.21)):

$$\left.\begin{aligned} P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0, 0) \mid Y_v = y_b) &= 1 \\ P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 0, 0) \mid Y_v \neq y_b) &= 0 \end{aligned}\right\}, \quad b = 1,...,B \ , \ v = 1,...,n. \tag{6.24}$$

Het voorlaatste element is nu 0, omdat Y_v als niet waargenomen wordt beschouwd. Aan (6.24) is eenvoudig in te zien dat in dit geval niet voldaan is aan de MAR-voorwaarde (6.10) om de designvariabele te negeren, immers de verdeling van de missing data indicator hangt af van niet-waargenomen variabelen. In groepsgerichte designs zijn we dus verplicht de achtergrondvariabele mee te nemen in de analyse. Zouden we dat niet doen dan geeft een MML-analyse wel uitkomsten, deze zijn echter onjuist. Met een voorbeeld zullen wij dit illustreren.

We genereren onder het Raschmodel itemantwoorden voor twee groepen van 500 leerlingen. De eerste groep van 500 minder vaardige personen, met waarde y_1 van de achtergrond-variabele, is aselect getrokken uit een normale verdeling met gemiddelde -1 en variantie 1, $N(-1,1)$. De tweede vaardiger groep, met de waarde y_2, is aselect getrokken uit $N(1,1)$. Voor de eerste groep worden itemantwoorden op vijf items die

gemakkelijk zijn ($\beta_i = -2, i = 1,...,5$) en vijf middelmatig moeilijke items ($\beta_i = 0, i = 6,...,10$) gegenereerd. De tweede groep maakt naast de middelmatig moeilijke items 6 tot en met 10, vijf items moeilijke items met $\beta_i = 2, i = 11,\ldots,15$. Voor de aldus gegenereerde antwoorden voeren we twee MML-analyses uit: in de eerste negeren we de achtergrond-variabele, in de tweede nemen we de achtergrondvariabele mee in de analyse. Het resultaat, waarbij de normering zodanig is gekozen dat $\sum_{i=1}^{15} \hat{\beta}_i = 0$, staat in tabel 6.4. We zien in tabel 6.4 dat het niet meenemen van de achtergrondvariabele in groepsgerichte designs systematisch verkeerde schattingen van de itemparameters oplevert. De gemakkelijke items 1 tot en met 5 worden moeilijker geschat dan ze in werkelijkheid zijn. Van de moeilijke items 11 tot en 15 worden itemparameter onderschat. Ook de parameters van de vaardigheids-verdeling, zie onder in de tabel, worden als gevolg van de gemaakte specificatiefout verkeerd geschat. Zoals in tabel 6.4 te zien zijn de afwijkingen van de ingevoerde parameters doorgaans meer dan 2 standaardfouten. Als we de achtergrondinformatie wel meenemen zien we dat zowel de itemparameters als de parameters van de vaardigheidsverdelingen, rekening houdend met de standaardfouten naar verwachting worden teruggeschat.

Tabel 6.4
MML-analyse gesimuleerd groepsgericht design

		negeren y_b		meenemen y_b	
item	β_i	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
1	-2	-1.847	.127	-2.158	.113
2	-2	-1.786	.127	-2.099	.112
3	-2	-1.726	.126	-2.042	.111
4	-2	-1.761	.126	-2.076	.112
5	-2	-1.679	.125	-1.996	.110
6	0	0.018	.074	0.006	.076
7	0	-0.003	.074	-0.016	.076
8	0	-0.036	.074	-0.050	.076
9	0	0.018	.074	0.006	.076
10	0	0.018	.074	0.006	.076
11	2	1.706	.125	2.035	.111
12	2	1.753	.126	2.080	.112
13	2	1.813	.127	2.139	.113
14	2	1.637	.125	1.967	.110
15	2	1.874	.127	2.198	.114
		$\hat{\mu}$ = 0.018(.083)	$\hat{\sigma}$ = 1.326(.053)	$\hat{\mu}_1$ = -0.984(.061)	$\hat{\sigma}_1$ = 0.980(.049)
				$\hat{\mu}_2$ = 1.018(.065)	$\hat{\sigma}_2$ = 1.062(.050)

Bij groepsgerichte designs moeten we dus in een MML-analyse de achtergrondvariabele meenemen en tegelijk met de itemparameters de verdelingsparameters van alle groepen meeschatten. Omdat standaardprogrammatuur voor MML, zoals BILOG (Mislevy & Bock, 1986), deze optie niet kent en suggereert dat het geen rol speelt moet men in de praktijk hiervoor op zijn hoede zijn.

6.5.2 CML in stochastische designs

In paragraaf 6.5 werd reeds opgemerkt dat Rubins voorwaarden niet beslissend zijn voor het eventueel negeren van de designvariabele in de CML-analyse. Alvorens de

mogelijkheden voor CML-analyse in de drie stochastische designvormen te bespreken, zullen we de reden hiervoor uiteenzetten en de voor CML beslissende voorwaarden formuleren.

Stel dat we gebruik zouden willen maken van Rubins 'ignorability' principe in een CML-analyse. Dan analyseren we uiteindelijk de marginale verdeling van de geobserveerde itemantwoorden (zie (6.9)):

$$\prod_{v=1}^{n} \int_{\boldsymbol{u}_{mis,v}} f_\tau(\boldsymbol{u}_{obs,v}, \boldsymbol{u}_{mis,v})\, d\boldsymbol{u}_{mis,v} = \prod_{v=1}^{n} P_{\beta,\theta_v}(\boldsymbol{x}_{obs,v}).$$

De verdeling van het geobserveerde antwoordpatroon $\boldsymbol{X}_{obs,v}$ hangt hierin af van de moeilijkheidsparameters β en de individuele vaardigheidsparameter θ_v, die bij CML in tegenstelling tot bij MML niet als toevalsvariabele wordt beschouwd. Om de CML-methode te kunnen toepassen zou er voor elke persoon v een voldoende steekproefgrootheid of statistiek $S_{obs,v} = S_{obs,v}(\boldsymbol{X}_{obs,v})$ moeten bestaan voor θ_v waarop we dan zouden kunnen conditioneren, zodat de aannemelijkheidsfunctie onafhankelijk van θ_v wordt. In onvolledige designs bestaat zo'n voldoende statistiek echter niet in de verdeling van $\boldsymbol{X}_{obs,v}$, hetgeen we nu aan de hand van een voorbeeld zullen illustreren.

Stel we hebben drie items die het Raschmodel volgen en we hebben een gerandomiseerd design met twee boekjes, bestaande uit respectievelijk item 1 en 2, en item 1 en 3. De verdeling van de itemindicator wordt gegeven door

$$P(\boldsymbol{R} = \boldsymbol{r}_1 = (1,1,0) = \phi, \quad \text{en} \quad P(\boldsymbol{R} = \boldsymbol{r}_2 = (1,0,1)) = 1 - \phi.$$

In het Raschmodel verwachten we, zie hoofdstuk 4, dat de somscore op de geobserveerde items

$$S_{obs,v} = \sum_{j \in obs,v} X_{vj}, \tag{6.25}$$

voldoende zal zijn voor θ_v en dat dus door conditioneren hierop er per persoon een voorwaardelijke kans geldt die alleen afhangt van de itemparameters. De somscore (6.25) is echter niet voldoende in de verdeling van $\boldsymbol{X}_{obs,v}$.

Merk allereerst op dat in het voorbeeld dat we bespreken de verdeling van $\boldsymbol{X}_{obs,v}$ en de verdeling van alle toevalsvariabelen $(\boldsymbol{X}_{obs,v}, \boldsymbol{R}_v)$ exact gelijk zijn. Er geldt namelijk altijd dat

$$P(\boldsymbol{x}_{obs,v}) = P(\boldsymbol{x}_{obs,v} \mid \boldsymbol{R}_v = \boldsymbol{r}_1) \,.\, P(\boldsymbol{R}_v = \boldsymbol{r}_1) + P(\boldsymbol{x}_{obs,v} \mid \boldsymbol{R}_v = \boldsymbol{r}_2) \,.\, P(\boldsymbol{R}_v = \boldsymbol{r}_2). \tag{6.26}$$

En voor de verdeling van $(\boldsymbol{X}_{obs,v}, \boldsymbol{R}_v)$ geldt

$$P(\boldsymbol{x}_{obs,v}, \boldsymbol{R}_v = \boldsymbol{r}_b) = P(\boldsymbol{x}_{obs,v} \mid \boldsymbol{R}_v = \boldsymbol{r}_b) \,.\, P(\boldsymbol{R}_v = \boldsymbol{r}_b) \text{ voor } b = 1,2. \qquad (6.27)$$

Als we nu kijken naar de mogelijke waarden van $\boldsymbol{X}_{obs,v}$, dan is dat of de waarneming $\{X_1 = x_1, X_2 = x_2\}$ of $\{X_1 = x_1, X_3 = x_3\}$. In het eerste geval is het tweede deel van het rechterlid van (6.26) gelijk aan 0 omdat $P(X_1 = x_1, X_2 = x_2 \mid \boldsymbol{r}_2 = (1,0,1)) = 0$; de kans op een antwoord op item 1 en 2, gegeven dat item 1 en 3 zijn waargenomen is immers 0. Verder volgt dan direct dat formule (6.26) in dat geval gelijk is met (6.27). In het tweede geval is, volgens dezelfde redenering, het eerste deel van het rechterlid gelijk aan 0 en ook (6.26) weer gelijk aan (6.27).

In ons voorbeeld gaan we, om een kortere notatie te krijgen, de itemparameters en de persoonsparameters transformeren, respectievelijk $\varepsilon_i = \exp(-\beta_i)$, $i = 1,2,3$ en $\exp(\theta) = \xi$. Vervolgens beschouwen we alle mogelijke uitkomsten waarvoor de somscore (6.25) gelijk aan 1 is en geven in tabel 6.5 de relevante kansen.

Tabel 6.5

Kansen op alle uitkomsten met $S_{obs} = 1$ in Raschmodel met drie items

$\boldsymbol{x}_{obs}, \boldsymbol{r}$	$P(\boldsymbol{x}_{obs}) = P(\boldsymbol{x}_{obs}, \boldsymbol{r})$	$P(\boldsymbol{x}_{obs} \mid \boldsymbol{r}_1)$	$P(\boldsymbol{x}_{obs} \mid \boldsymbol{r}_2)$
(1)	(2)	(3)	(4)
$x_1=1, x_2=0, 110$	$\dfrac{\phi.\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	$\dfrac{\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	0
$x_1=0, x_2=1, 110$	$\dfrac{\xi\varepsilon_2}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	$\dfrac{\xi\varepsilon_2}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	0
$x_1=1, x_3=0, 101$	$\dfrac{(1-\phi).\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$	0	$\dfrac{\xi\varepsilon_1}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$
$x_1=0, x_3=1, 101$	$\dfrac{(1-\phi).\xi\varepsilon_3}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$	0	$\dfrac{\xi\varepsilon_3}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$
1	$\dfrac{\phi.\xi(\varepsilon_1+\varepsilon_2)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)} + \dfrac{(1-\phi).\xi(\varepsilon_1+\varepsilon_3)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$	$\dfrac{\xi(\varepsilon_1+\varepsilon_2)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_2)}$	$\dfrac{\xi(\varepsilon_1+\varepsilon_3)}{(1+\xi\varepsilon_1)(1+\xi\varepsilon_3)}$
s_{obs}	$P(s_{obs})$	$P(s_{obs} \mid \boldsymbol{r}_1)$	$P(s_{obs} \mid \boldsymbol{r}_2)$

In kolom (1) van tabel 6.5 staan alle mogelijke uitkomsten. Beschouwen we eerst kolom (2). Hierin staan in het bovenste deel de kansen op deze uitkomsten en in het onderste deel de kans dat $S_{obs} = 1$. De voorwaardelijk kans op een willekeurige uitkomst,

gegeven $s_{obs} = 1$, verkrijgen we door het delen van de term uit het onderste deel van de tabel door een term uit het bovenste deel. Er geldt immers

$$P(\boldsymbol{x}_{obs}, \boldsymbol{r}) = \frac{P(\boldsymbol{x}_{obs}, \boldsymbol{r}, s_{obs})}{P(s_{obs})} = \frac{P(\boldsymbol{x}_{obs}, \boldsymbol{r})}{P(s_{obs})}.$$

Als we zo'n deling uitvoeren zien we dat het resultaat afhangt van individuele parameter ξ. Waaruit volgt dat S_{obs} niet voldoende is voor ξ en dus ook niet voor θ, en we kunnen CML dus niet toepassen in de verdeling van $\boldsymbol{X}_{obs}$ of van $(\boldsymbol{X}_{obs}, \boldsymbol{R})$.

Wat er echter wel mogelijk is zien we in de kolommen (3) en (4) van tabel 6.5. Hierin staan voor ons voorbeeld de conditionele kansen op de uitkomsten, $P(\boldsymbol{x}_{obs} \mid \boldsymbol{R}_v = \boldsymbol{r}_b)$, $b = 1,2$, een de conditionele kans dat de somscore 1 is, $P(S_{obs} = 1 \mid \boldsymbol{R}_v = \boldsymbol{r}_b)$, $b = 1,2$, beiden gegeven de waarde van itemindicator variabele. Eenvoudig is na te gaan dat in de conditionele verdeling van $\boldsymbol{X}_{obs}$ gegeven $\boldsymbol{R}$ de somscore wel voldoende is voor de individuele parameter ξ. De kans op een uitkomst gegeven de somscore bepalen we in deze conditionele verdelingen weer door in tabel 6.5 de kans uit het onderste deel te delen op een term uit het bovenste deel. Er geldt namelijk:

$$\frac{P(\boldsymbol{x}_{obs} \mid \boldsymbol{r})}{P(s_{obs} \mid \boldsymbol{r})} = \frac{P(\boldsymbol{x}_{obs}, s_{obs} \mid \boldsymbol{r})}{P(s_{obs} \mid \boldsymbol{r})} = P(\boldsymbol{x}_{obs} \mid s_{obs}, \boldsymbol{r}). \tag{6.28}$$

Voor alle gegeven uitkomsten en ook voor de andere uitkomsten is eenvoudig na te gaan dat het resultaat van deze deling onafhankelijk is van de individuele parameter ξ.

In de conditionele verdelingen, gegeven de itemindicator, zitten we dus in dezelfde positie als in het Raschmodel voor volledige data: we hebben een voldoende statistiek waarmee voor elke persoon de individuele parameter kunnen uitconditioneren uit de aannemelijkheidsfunctie. Daarmee is dan ook voldaan aan de eerste voorwaarde om de CML-schattingsmethode te kunnen toepassen. Merk op aan (6.28) dat we alternatief zouden kunnen zeggen dat alleen S_{obs} en $\boldsymbol{R}$ gezamenlijk voldoende zijn voor de individuele parameter ξ of θ. Ging het in de theorie van Rubin (1976) en ook in paragraaf 6.5.1, waar we MML in stochastische designs bespraken, steeds om de vraag of we in de analyse de designvariabele konden negeren, bij CML is deze vraag niet aan de orde. Willen we CML toepassen dan zullen we de designvariabele expliciet in de analyse moeten meenemen, omdat er anders geen voldoende statistiek voor de individuele vaardigheid bestaat. Dus Rubins voorwaarden kunnen niet beslissend zijn

voor de toepassing van CML in stochastische onvolledige designs. Welke dat wel zijn gaan we nu behandelen.

Als we CML gaan toepassen gaan we dus uit van de verdeling van alle waargenomen toevalsvariabelen. In het algemeen kan dit geschreven worden als:

$$P_{\theta,\beta,\phi}(\boldsymbol{x_{obs}}, \boldsymbol{r}) = \prod_{v=1}^{n} P_{\theta_v,\beta,\phi}(\boldsymbol{x}_{obs,v} \mid \boldsymbol{r}_v) \cdot P_{\phi}(\boldsymbol{r}_v). \tag{6.29}$$

We gebruiken dezelfde notatie als eerder. We onderscheiden B waarden van de design-variabele $\boldsymbol{r}_b$, $b = 1,\ldots,B$; n_b is het aantal personen dat boekje b maakt; $\beta_{(b)}$ is de k_b-vector met de parameters van de items in boekje b. Dan kunnen we (6.29) herschrijven als:

$$P_{\theta,\beta,\phi}(\boldsymbol{x_{obs}}, \boldsymbol{r}) = \prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\theta_v,\beta_{(b)},\phi}(\boldsymbol{x}_{obs,v} \mid \boldsymbol{r}_b) \cdot \prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b). \tag{6.30}$$

We zien in (6.30) dat we de aannemelijkheidsfunctie van alle waarnemingen kunnen schrijven als het produkt van twee termen. Het is in te zien dat het eerste deel van het rechterlid van (6.30) niets anders is dan het produkt van B volledige data aannemelijkheidsfuncties, zoals in hoofdstuk 4 is besproken. In elk boekje is er, zoals bij de volledige data, zoals we in het voorgaande zagen (6.28), voor elke persoon een voldoende statistiek S_{obs}, zodat geldt

$$\prod_{v=1}^{n_b} P_{\theta_v,\beta_{(b)},\phi}(\boldsymbol{x}_{obs,v} \mid \boldsymbol{r}_b) = \prod_{v=1}^{n_b} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid s_{obs,v}, \boldsymbol{r}_b) \cdot P_{\theta_v,\beta,\phi}(s_{obs,v} \mid \boldsymbol{r}_b). \tag{6.31}$$

Het eerste deel van het rechterlid van (6.31) hangt alleen nog maar af van de itemparameters $\beta_{(b)}$ en dit deel wordt in de CML-methode gemaximaliseerd naar de parameters β in plaats van het linkerlid. De maxima geven de itemparameterschattingen. De rechtvaardiging van de CML-methode hangt mede af van het feit of we het tweede deel van het rechterlid van (6.31) mogen weglaten uit de analyse. Zou het tweede deel van het rechterlid onafhankelijk zijn van β dan is het duidelijk dat het niet uitmaakt of we het linkerlid, de volledige aannemelijkheids-functie, dan wel alleen het eerste deel van het rechterlid, de conditionele aannemelijkheids-functie gebruiken. We zien echter dat ook het tweede deel van het rechterlid van (6.31), de verdeling van S_{obs}, afhangt van β. Het zo maar weglaten van dit deel zal in zijn algemeenheid natuurlijk niet dezelfde resultaten voor de itemparameterschattingen opleveren. Het is echter aangetoond (Andersen, 1973b) dat voor IRT-modellen die behoren tot de exponentiële familie, zie hoofdstuk 4, zoals het

Raschmodel en het OPLM model, die afhankelijkheid van het tweede lid van β een zeer speciale structuur heeft, waardoor het in dat geval gerechtvaardigd is het in de analyse buiten beschouwing te laten, en dat de resulterende schattingen de in hoofdstuk 4 gememoreerde goede statistische eigenschappen hebben. De speciale structuur komt er op neer dat de verdeling van S_{obs} niet rechtstreeks afhankelijk is van β; de afhankelijkheid is altijd gekoppeld aan de afhankelijkheid van de persoonsparameter. We zullen hier niet verder op ingaan en verwijzen voor details naar Andersen (1973b).

De voorgaande beschouwing geldt voor elk volledig boekje in onvolledige designs en natuurlijk ook voor aannemelijkheidsfunctie voor alle boekjes. Dus het is in onze modellen gerechtvaardigd om ook in onvolledige designs in plaats van het produkt over B boekjes van het linkerlid van (6.31) uit te gaan van het produkt over B boekjes van het eerste deel van het rechterlid: de conditionele aannemelijkheidsfunctie:

$$L_c = \prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\beta_{(b)}}(\boldsymbol{x}_{obs,v} \mid s_{obs,v}, \boldsymbol{r}_b) \tag{6.32}$$

Of het in stochastische designs gerechtvaardigd is om alleen (6.32) te beschouwen, hangt dan alleen nog maar af van de vraag of we ook het rechterdeel van de aannemelijkheidsfunctie (6.30):

$$\prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b), \tag{6.33}$$

in de analyse weg kunnen laten. Het antwoord hierop is analoog aan de redenering hiervoor. Zolang (6.33) onafhankelijk is van de itemparameters β, dan is dat gerechtvaardigd. Als er afhankelijkheid is dan moet voor de rechtvaardiging van CML in stochastische designs de eerder omschreven speciale structuur aanwezig zijn. Is er rechtstreekse afhankelijkheid van (sommige) itemparameters in (6.33) dan is CML niet toegestaan. We bespreken nu de mogelijkheid van CML voor de drie stochastische designvormen.

CML in gerandomiseerde onvolledige designs

De designverdeling in gerandomiseerde designs wordt gegeven door (6.2):

$$\prod_{b=1}^{B} \prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b) = \prod_{b=1}^{B} \prod_{v=1}^{n_b} \phi_b. \tag{6.34}$$

En we zien dat (6.34) geheel onafhankelijk is van de itemparameters β, en dus dat toepassen van CML in gerandomiseerde onvolledige designs evenals bij MML geen problemen oplevert.

CML in meerfasen onvolledige designs

In meerfasen onvolledige designs kunnen we (6.33), met behulp van de itemindicator verdeling (6.4), schrijven als:

$$\prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b) = \prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b \mid \boldsymbol{x}_{obs,v}).P_{\beta_{(obs)},\theta_v}(\boldsymbol{x}_{obs,v}). \qquad (6.35)$$

In (6.35) zien we dat het tweede deel van het rechterlid rechtstreeks afhangt van de itemparameters van de items, waarvan de waargenomen waarden bepalen wie welk boekjes gaat maken. De speciale afhankelijkheidsstructuur, waarvan bij de rechtvaardiging van CML in het algemeen sprake is, is hier niet aanwezig. CML in meerfasen designs is dus niet mogelijk. Dit in tegenstelling tot MML waarbij, zoals we eerder zagen in paragraaf 6.5.1, de designvariabele in de analyse kon worden genegeerd om tot correcte resultaten te komen. Wij zullen dit met een voorbeeld met gesimuleerde data illustreren. Daarvoor beschouwen opnieuw het voorbeeld uit paragraaf 6.3.2. De tien middelmatig moeilijke items 1 tot 10 uit de sorteertoets hebben een moeilijkheid in het Raschmodel van 0. Voor de gemakkelijke items is $\beta_i = -1, i = 11,...,15$ en voor de moeilijke $\beta_i = 1\ , i = 16,...,20$. Als we 1000 itemantwoorden genereren voor vaardigheden getrokken uit een standaard normale verdeling en in de analyse alleen de antwoorden op de moeilijke items beschouwen voor de personen met een score van 6 of meer op de sorteertoets en de antwoorden op de gemakkelijke items alleen voor de personen met een score van 5 of minder op de sorteertoets, dan leveren analyses van deze gegevens de resultaten op uit tabel 6.6.

We zien in tabel 6.6 dat in de MML-analyse de itemmoeilijkheden bij het negeren van de designvariabele in dit tweefasen design goed worden geschat: er zijn geen geschatte moeilijkheden $\hat{\beta}_i$ die meer dan twee geschatte standaardfouten van de ingevoerde moeilijk-heden afliggen. Hetzelfde geldt voor de verdelingsparameters die onder in de tabel staan vermeld. Voor de CML-schattingen van de moeilijkheid geldt dit alleen maar voor de items van de sorteertoets (1 tot 10). Ze verschillen nauwelijks van de MML-schattingen. De overige itemmoeilijkheden worden systematisch onjuist geschat. De gemakkelijke items (11 tot 15) worden gemakkelijker geschat dan ze in werkelijkheid zijn en de moeilijke items (16 tot 20) moeilijker. Steeds is het verschil

tussen de geschatte moeilijkheid $\hat{\beta}_j$ en de echte moeilijkheid β_j meer dan twee geschatte standaardfouten. Tenslotte zij opgemerkt dat in de realisatie van deze simulatie van de 1000 personen die de sorteertoets maakten er vervolgens 556 met de gemakkelijke items verder gingen en 444 met de moeilijke. Dit verklaart de verschillen tussen de items in de geschatte standaardfouten in tabel 6.6.

Tabel 6.6

CML- en MML-analyse gesimuleerd meerfasen design

		CML		MML	
Item	β_i	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
1	0	0.043	.069	0.064	.068
2	0	-0.101	.069	-0.075	.069
3	0	-0.007	.069	0.016	.068
4	0	-0.081	.069	-0.056	.069
5	0	-0.036	.069	-0.013	.068
6	0	-0.076	.069	-0.051	.069
7	0	0.038	.069	0.059	.068
8	0	0.023	.069	0.044	.068
9	0	-0.026	.069	-0.003	.068
10	0	-0.071	.069	-0.046	.069
11	-1	-1.391	.090	-1.144	.097
12	-1	-1.286	.089	-1.033	.095
13	-1	-1.192	.090	-0.933	.095
14	-1	-1.310	.090	-1.058	.096
15	-1	-1.318	.090	-1.067	.096
16	1	1.314	.098	1.012	.105
17	1	1.410	.099	1.114	.106
18	1	1.420	.099	1.124	.106
19	1	1.381	.098	1.083	.106
20	1	1.266	.098	0.962	.105
	$\mu = 0$			$\hat{\mu} = 0.026(.038)$	
	$\sigma = 1$			$\hat{\sigma} = 0.944(.031)$	

Uit dit voorbeeld moge duidelijk zijn dat CML in een meerfasen design geen correcte resultaten oplevert en dus niet toegestaan is. Aangezien standaardprogrammatuur voor CML-analyse, bijvoorbeeld OPLM, geen rekening houdt met hoe de onvolledige gegevens zijn ontstaan, dient men hiervoor op de hoede te zijn.

CML in groepsgerichte designs

In groepsgerichte designs is (6.33) af te leiden uit de verdeling van de itemindicator variabele (6.5):

$$\prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\phi}(\boldsymbol{R}_v = \boldsymbol{r}_b) = \prod_{b=1}^{B}\prod_{v=1}^{n_b} P_{\pi_b}(Y_v = y_b). \qquad (6.36)$$

Het zal duidelijk zijn dat uitdrukking (6.36) niet van de itemparameters β afhangt. De kans dat een persoon tot een bepaalde groep b behoort wordt natuurlijk niet bepaald door de items die deze persoon maakt. Hieruit volgt dat CML met de conditionele aannemelijkheidsfunctie (6.32) in groepsgerichte stochastische designs zonder problemen kan plaatsvinden.

Ter illustratie volgt tenslotte het resultaat van de CML-analyse van de gesimuleerde gegevens in een groepsgericht design, waarvoor in tabel 6.4 de resultaten van de MML-analyses werden gegeven.

Tabel 6.7
CML-analyse in een gesimuleerd groepsgericht design

item	β_i	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
1	-2	-2.158	.113
2	-2	-2.099	.112
3	-2	-2.042	.111
4	-2	-2.076	.112
5	-2	-1.996	.110
6	0	0.006	.076
7	0	-0.016	.076
8	0	-0.050	.076
9	0	0.006	.076
10	0	0.006	.076
11	2	2.035	.111
12	2	2.080	.112
13	2	2.139	.113
14	2	1.967	.110
15	2	2.198	.114

In tabel 6.7 zien we dat alle CML-schattingen van de moeilijkheid $\hat{\beta}_i$ in dit groepsgerichte design minder dan twee standaardfouten van de ingevoerde waarden β_i afliggen. Als we resultaten vergelijken met de MML-analyse, waarbij we de achtergrondvariabele Y expliciet in de analyse meenemen, zie tabel 6.4, dan zien dat resultaten bijna perfect overeenstemmen.

De omstandigheid dat CML-analyses zelfs in stochastische groepsgerichte designs zonder problemen kunnen worden uitgevoerd is nog eens bevestiging van het feit, dat bij CML, ook bij volledige designs, geen rekening gehouden hoeft te worden met de wijze waarop de steekproef personen uit een populatie is getrokken. Dit in tegenstelling tot MML, waarbij altijd expliciet rekening moet worden gehouden met de wijze van steekproeftrekking en met het in dit geval relevante lidmaatschap van subpopulaties van personen.

6.6 Schatten van persoonsparameters in stochastische designs

Voor de persoonsparameterschattingen zijn in de IRT verschillende methoden beschikbaar. In paragraaf 4.5 werden behandeld de ML-schatter (grootste aannemelijkheid), de WML- schatter (gewogen-grootste-aannemelijkheid) en de EAP-schatter (de verwachting van de a posteriori verdeling van de vaardigheid). Bij het schatten van de persoonsparameter θ_v gaan we ervan uit dat de itemparameters uit het IRT-model waar we mee werken voldoende nauwkeurig zijn geschat om ze bekend te veronderstellen. We gaan dus uit van gecalibreerde itemverzamelingen. Reeds in paragraaf 6.1 werd gesteld dat een van de positieve eigenschappen van het werken met IRT-modellen is dat de vaardigheid van de personen met verschillende opgaven, deelverzamelingen uit een gecalibreerde itemverzameling, op dezelfde schaal worden geschat. Deze eigenschap impliceert dat voor de schatting van de vaardigheid de designvariabele geen rol speelt in de analyse. In deze paragraaf zullen nagaan of dit in het algemeen bij de drie besproken stochastische designtypen ook het geval is. We moeten daarbij in de bespreking onderscheid maken naar enerzijds de ML- en de WML-schatter en anderzijds de EAP-schatter van θ_v.

6.6.1 ML- en WML-vaardigheidsschatting in stochastische designs

In stochastische designs is steeds de vraag aan de orde of we in de analyse rekening moeten houden met het toevalsproces dat de designs genereert, dan wel dat we het stochastisch karakter van de designvariabele kunnen negeren. Omdat in de ML-schatting en de WML-schatting van de persoonsparameter dezelfde toevalsvariabele wordt beschouwd, namelijk het antwoordpatroon van persoon v op de items $\boldsymbol{X}_v = (X_{v1},...,X_{vk})$, heeft deze vraag bij beide methoden hetzelfde antwoord. We zullen daarom alleen de ML schatting nader beschouwen. De theorie van Rubin, behandeld in paragraaf 6.5. is ook hier weer direct toepasbaar.

In de eerdere notatie is de toevalsvariabele die ons interesseert $\boldsymbol{U}_v = \boldsymbol{X}_v$ waarvan de verdeling $f_\tau(\boldsymbol{u}_v)$ alleen afhangt van de onbekende parameter $\tau = \theta_v$. In gerandomiseerde en in meerfasen designs deelt de missing data indicator $\boldsymbol{M}_v$, die hier hetzelfde is als de itemindicator $\boldsymbol{R}_v$, de variabelen $\boldsymbol{U}_v$ op in:

$$\boldsymbol{U}_{obs,v} = \boldsymbol{X}_{obs,v} \text{ en } \boldsymbol{U}_{mis,v} = \boldsymbol{X}_{mis,v}.$$

In deze gevallen is eenvoudig na te gaan dat de verdeling van de itemindicator, respectievelijk (6.2) voor gerandomiseerde design en (6.4) voor meerfasen designs, op zijn minst voldoet aan de MAR-voorwaarde (6.10) voor het negeren van het design in

de analyse. Dus in deze designs kan de schatting gebaseerd worden op de marginale verdeling van de observaties:
$P_{\theta_v}(\boldsymbol{x}_{obs,v})$. Opgemerkt kan worden dat het negeren van de designvariabele bij het schatten van de persoonsparameter eveneens gerechtvaardigd is bij het adaptief toetsen, hetgeen immers een limietgeval is van meerfasen toetsen (zie paragraaf 6.3.2).

Bij groepsgerichte designs moet bij het schatten van de persoonsparameter analoog bij de MML-calibratie (paragraaf 6.3.3) onderscheid gemaakt worden tussen het wel en niet meenemen van de achtergrondvariabele Y in de analyse. Bij wel meenemen geldt

$$\boldsymbol{U}_{obs,v} = (\boldsymbol{X}_{obs,v}, Y_v) \text{ en } \boldsymbol{U}_{mis,v} = \boldsymbol{X}_{mis,v}. \tag{6.37}$$

De verdeling van de missing data indicator is (vergelijk met (6.21)):

$$\left.\begin{array}{l} P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1) \mid Y_v = y_b) = 1 \\ P(\boldsymbol{M}_v = (\boldsymbol{r}_b, 1) \mid Y_v \neq y_b) = 0 \end{array}\right\}, \quad b = 1,...,B; v = 1,...,n. \tag{6.38}$$

In (6.38) is $\boldsymbol{r}_b$ weer de $k-$vector met k_b maal een 1 op plaatsen die de geobserveerde items in boekje b indiceren, en $k-k_b$ maal een 0. De laatste 1 in de waarde van $\boldsymbol{M}_v$ indiceert het waarnemen van Y_v. Duidelijk is dan dat aan de MAR-voorwaarde (6.10) is voldaan en we in de analyse de designvariabele kunnen negeren en ons kunnen baseren op de marginale verdeling van de observaties $P_{\theta_v,\pi_b}(\boldsymbol{x}_{obs,v}, y_v)$. Merk op dat we deze verdeling kunnen schrijven als:

$$P_{\theta_v,\pi_b}(\boldsymbol{x}_{obs,v}, y_v) = P_{\theta_v}(\boldsymbol{x}_{obs,v} \mid y_v) \, . \, P_{\pi_b}(Y_v = y_b). \tag{6.39}$$

In (6.39) zien we dat voor het maximaliseren ervan naar θ_v we kunnen volstaan met het maximaliseren van het eerste deel van het rechterlid. In de IRT-modellen die wij beschouwen geldt hiervoor, vanwege de lokale stochastische onafhankelijkheid:

$$P_{\theta_v}(\boldsymbol{x}_{obs,v} \mid y_v) = \prod_{j \in obs,v} P_{\theta_v}(x_{vj} \mid y_v) = \prod_{j \in obs,v} P_{\theta_v}(x_{vj}). \tag{6.40}$$

Hierin staat $P_{\theta_v}(x_{vj})$ voor het IRT-model dat we beschouwen. We zien dus dat de aannemelijkheidsfunctie (6.40) die we, eventueel vermenigvuldigd met een functie van θ bij WML, die we maximaliseren voor het verkrijgen van de persoonsparameterschatting onafhankelijk is van de achtergrondvariabele Y. Dus ook hier geldt dat de persoons-

parameterschatting onafhankelijk is van de toevallige items, hier bepaald door de waarde van de achtergrondvariabele, die uit de gecalibreerde itemverzameling zijn afgenomen.

Als we in groepsgerichte designs de achtergrondvariabele niet zouden meenemen dan krijgen we voor de opdeling door de designvariabele van alle variabelen in plaats van (6.37):

$$\boldsymbol{U}_{obs,v} = \boldsymbol{X}_{obs,v} \text{ en } \boldsymbol{U}_{mis,v} = (\boldsymbol{X}_{mis,v}, Y_v). \tag{6.41}$$

En de verdeling van de designvariabele is als in (6.38), met dien verstande dat het laatste element altijd de waarde 0 heeft in plaats van 1, welke niet voldoet aan de MAR-voorwaarde (6.10), hetgeen betekent dat het design niet genegeerd kan worden. In dit geval echter zou het negeren geen consequenties hebben: het alleen beschouwen van de marginale verdeling van de observaties $P_{\theta_v}(\boldsymbol{x}_{obs,v})$ levert, vanwege eigenschap (6.40), dezelfde uitdrukking op voor de aannemelijkheidsfunctie als bij het wel meenemen van de achtergrondvariabele.

6.6.2 EAP vaardigheidsschatting in stochastische onvolledige designs

De EAP-schatter voor de vaardigheid is in tegenstelling tot alle voorgaande schattingsmethoden een bayesiaanse schatter en geen grootste-aannemelijkheidsschatter. Dat betekent dat de algemene theorie voor het negeren van de designvariabele in de analyse, zoals behandeld in paragraaf 6.5, hier niet direct van toepassing is. Rubin (1976) heeft echter ook voor bayesiaanse schattingsmethoden aangegeven onder welke voorwaarden het design in de analyse genegeerd kan worden. Het zou in het kader van dit boek te ver voeren om ook dit onderwerp uitgebreid te behandelen. We volstaan met op te merken dat voor het negeren van het design in een bayesiaanse analyse naast de voorwaarden die al gelden voor de ML-schattingen nog een extra voorwaarde moet gelden. Of aan deze voorwaarde voldaan is zullen we hierna voor de drie besproken stochastische designtypen kort bespreken.

De extra voorwaarden heeft betrekking op de eigenschappen van de a priori verdelingen die in de bayesiaanse analyse worden gebruikt. In het algemeen is aan de voorwaarden voor het negeren van de designvariabele in een bayesiaanse analyse voldaan, als de a priori verdelingen van de betrokken parameters onafhankelijk zijn. Bij het schatten van de persoonsparameters in stochastische designs hebben we te maken met twee parameters: de persoonsparameter θ en de parameter ϕ van de

verdeling van de designvariabele. Bij de mogelijkheid de designvariabele te negeren bij de EAP-schatting van θ zullen we de a priori relatie tussen deze parameters moeten beschouwen.

In gerandomiseerde designs zal er geen enkele a priori relatie zijn tussen θ en ϕ. Voor de gezamenlijke a priori verdeling van deze parameters zal dan ook voldaan zijn aan de onafhankelijkheidsvoorwaarde:

$$P(\theta,\phi) = P(\theta) \,.\, P(\phi). \tag{6.42}$$

Omdat ook aan de MAR-voorwaarde is voldaan levert het negeren van het design ook voor de EAP-schatting van θ geen probleem op.

Hetzelfde geldt voor meerfasen designs: de parameter ϕ wordt volledig bepaald door uitkomsten van waargenomen variabele, die op zichzelf natuurlijk wel van de vaardigheid θ afhangen, maar voor de waarnemingen zijn gedaan is er geen enkele aanname over het verband tussen θ en ϕ. Dus ook hier is de aanname (6.42) reëel. Met het voldoen aan de MAR-voorwaarde is dit samen voldoende om ook in meerfasen designs bij het bepalen van de EAP-schatting de designvariabele in de analyse te negeren. Zowel bij gerandomiseerde als meerfasen designs kunnen we dus, na specificatie van een a priori verdeling, met behulp van (4.119) en (4.120) een EAP-schatting bepalen.

Anders is de situatie bij groepsgerichte designs daar hebben we al in paragraaf 6.6.1 al gezien dat om te voldoen aan de MAR-voorwaarde de achtergrondvariabele in de analyse moeten meenemen. Echter ook geredeneerd vanuit de a priori verdelingen is het in te zien dat het a priori aannemen van onafhankelijkheid van θ en ϕ hier niet reëel is. De parameter van de designverdeling ϕ wordt immers volledig bepaald door de achtergrondvariabele. Zouden we (6.42) aannemen dat zou dat betekenen dat we a priori geen relatie zien tussen de vaardigheid θ en de waarde van achtergrondvariabele Y, echter de relatie tussen deze twee variabelen is evenwel juist de reden om met groepsgerichte designs te werken. Dus (6.42) geldt zeker niet. Om toch EAP-schatters te kunnen verkrijgen in groepsgerichte designs zullen we dus Y expliciet in de analyse moeten meenemen. Om te voldoen aan Rubins voorwaarden hebben we de geldigheid van (6.42) niet meer nodig echter alleen dat er gegeven de achtergrondvariabele, onafhankelijkheid is tussen de a priori verdelingen:

$$P(\theta,\phi \mid Y_v = y_b) = P(\theta \mid Y_v = y_b) \,.\, P(\phi \mid Y_v = y_b).$$

Deze aanname omtrent de a priori verdeling van parameters zal in de praktijk geen problemen opleveren. Voor een persoon v in groepsgerichte designs, met waarde y_b van achtergrond-variabele, kan de EAP-schatting dan met a priori verdeling $g(\theta) = P(\theta \mid \boldsymbol{Y}_v = y_b)$ bepaald worden.