8

Equivaleren

Een leerling van het VWO doet een herexamen (tweede tijdvak) voor het vak natuurkunde en behaalt een hogere score dan tijdens het reguliere examen (eerste tijdvak). Waarom? We zouden kunnen concluderen dat deze hogere score een grotere vaardigheid weerspiegelt: de leerling heeft tussen de beide examens flink wat bijgeleerd. Aan de andere kant is het mogelijk dat het examen uit het tweede tijdvak gemakkelijker was dan dat uit het eerste. Zelfs bij een gelijk gebleven vaardigheid zou de leerling dan een hogere score behalen. Gezien het grote belang dat examens hebben, is het duidelijk dat de leerling een score moet krijgen die een zo goed mogelijke afspiegeling van haar of zijn vaardigheid is, ongeacht welk examen gemaakt is. Dit betekent in ieder geval dat voor iedere score op het tweede tijdvak een score op het eerste tijdvak gevonden moet worden die dezelfde vaardigheid representeert. Het zoeken van vergelijkbare scores is een voorbeeld van wat men equivaleren noemt.

De psychometrische theorie over equivaleren is zeer omvangrijk. Voor overzichten verwijzen we naar Angoff (1971), Holland en Rubin (1982) en Petersen, Kolen en Hoover (1989). In dit hoofdstuk zullen wij ons zoveel mogelijk beperken tot het behandelen van equivaleermethoden die in de praktijk veelvuldig gebruikt worden. De belangrijkste factor die bepalend is voor de wijze waarop de equivalering plaatsvindt is het gebruikte meetmodel. Zoals we gezien hebben in de hoofdstukken 3, 4 en 5 heeft elk model zijn eigen manier om met een toets de vaardigheid te bepalen. Voor de bepaling van de vaardigheid gebruiken we in de klassieke testtheorie (KTT) doorgaans geobserveerde scores op een toets, terwijl in de itemresponstheorie (IRT) de vaardigheid als parameter, die in het model is opgenomen, geschat wordt. Alvorens echter het equivaleren per meetmodel te bespreken, zullen we in paragraaf 8.1 eerst een globaal overzicht geven van het equivaleren. Aspecten die daarbij aan de orde zullen komen spelen zowel een rol bij equivaleren in de KTT als in de IRT. In paragraaf 8.2 gaan we vervolgens de equivalering in de KTT behandelen. In paragraaf 8.3 volgt equivaleren in de IRT. In de laatste paragraaf 8.4 worden de conclusies en aanbevelingen uit dit hoofdstuk kort samengevat.

8.1 Overzicht equivaleren

Zoals uit de inleiding blijkt, ontstaat de behoefte aan equivaleren als we de vaardigheid van twee personen met een verschillend meetinstrument meten en de resultaten met elkaar willen vergelijken. De eerste vraag die we hierbij zouden moeten beantwoorden is of equivaleren in de praktijk niet vermeden kan worden. Men zou kunnen denken dat in het voorbeeld uit de inleiding geen problemen waren ontstaan als het examen van het tweede tijdvak hetzelfde geweest was als dat van het eerste tijdvak. Omdat de examens identiek zijn, zullen ook de scores op beide examens gelijk dezelfde vaardigheid weerspiegelen. Het is maar al te duidelijk dat we niet op deze manier te werk kunnen gaan. Leerlingen die tijdens het tweede tijdvak examen doen, zijn dan bevoordeeld daar zij de inhoud van het af te nemen examen reeds kennen. Daarom, op grond van eerlijkheid, is het noodzakelijk om het herexamen verschillend van het eerste te laten zijn. Om de scores van een leerling, of meer algemeen voor verschillende leerlingen, op twee verschillende examens op een zinvolle manier met elkaar te kunnen vergelijken, zal men dus rekening moeten houden met de, mogelijk verschillende, moeilijkheid van beide examens. Het is immers onterecht als een tweede tijdvak kandidaat een hoger cijfer haalt dan een eerste tijdvak kandidaat, alleen maar omdat zij of hij een eenvoudiger examen gemaakt heeft.

Het ideaal van het vermijden van equivaleren wordt in zekere zin bereikt, zoals we later zullen zien, als we toetsen samenstellen uit een itembank die gecalibreerd is onder een IRT-model. In de praktijk is evenwel meestal het equivaleerprobleem aan de orde als we de scores op twee bestaande, vaste, toetsen vergelijkbaar willen maken. Overigens is in de KTT een andere werkwijze ook niet mogelijk, omdat we daar altijd uitgaan van de score op een toets. We zullen in dit hoofdstuk het equivaleerprobleem dan ook via deze weg benaderen.

Meer algemeen gesteld zouden we het probleem van het equivaleren als volgt kunnen omschrijven. Twee of meer groepen personen maken verschillende versies van een toets. Hoe kunnen de scores op de ene toets vertaald of naar een zelfde schaal getransformeerd worden als de scores op de andere toets, opdat ze vergelijkbaar worden? Het zal blijken dat het equivaleren van twee toetsen in feite neerkomt op het vinden van een functie die de scores op een toets Y transformeert naar de schaal van de scores op een toets X. Deze functie, die we de equivaleerfunctie noemen, noteren we met $e_X(Y)$. Het zal duidelijk zijn dat als we twee toetsen kunnen equivaleren, we ook meer toetsen kunnen equivaleren. In dit hoofdstuk zullen we dan ook steeds spreken over het equivaleren van twee toetsen.

We kunnen stellen dat de vergelijking van de scores op twee toetsen niet mag afhangen van wie welke toets heeft gemaakt. De score van een persoon op een toets zal echter afhangen van de moeilijkheid van de voorgelegde toets. Ook de twee situaties waarin de toetsen werden afgenomen mag op de vergelijking niet van invloed zijn. De score op een toets kan immers ook afhangen van externe factoren, zoals lawaai of extreme hitte tijdens de afname. Helaas zijn deze laatste effecten in de toetspraktijk vaak aanwezig. Alhoewel het soms mogelijk is om voor een lagere score tengevolge van externe factoren te corrigeren, zullen we ons hier in dit hoofdstuk niet mee bezig houden. Als we spreken over equivaleren dan willen we alleen corrigeren voor verschil in moeilijkheid.

In de praktijk kunnen we twee situaties onderscheiden waarin we willen equivaleren. In de eerste plaats is dat de situatie waarin we willen corrigeren voor niet geplande verschillen tussen de toetsen. Bij deze zogenaamde horizontale equivalering gaan we ervan uit dat we twee toetsen hebben die in principe hetzelfde meten, van dezelfde moeilijkheidsgraad zijn en bedoeld zijn voor één populatie. In deze situatie willen we dus onbedoelde ruis in metingen wegwerken. Deze ruis kan ontstaan doordat het bijvoorbeeld niet geheel gelukt is twee even moeilijke toetsen te maken. Het kan ook voorkomen dat de groepen leerlingen die de toetsen maken toch op de een of andere manier een weinig in vaardigheid verschillen. Een voorbeeld waar horizontale equivalering wordt toegepast is de Eindtoets Basisonderwijs van het Cito (in het vervolg Eindtoets). De Eindtoets, welke bestaat uit de drie onderdelen taal, rekenen en informatieverwerking, is een schoolvorderingentoets die jaarlijks wordt afgenomen in groep 8 van de basisschool. Deze toets heeft twee functies. Enerzijds levert de Eindtoets informatie over individuele leerlingen in verband met de overgang naar het voortgezet onderwijs, anderzijds levert de toets informatie ten behoeve van de evaluatie van het gegeven onderwijs (Uiterwijk & Engelen, 1993). Bij de constructie van een nieuwe versie van deze toets wordt er, onder andere, expliciet naar gestreefd om deze dezelfde moeilijkheidsgraad te geven als de oudere versie. Bovendien valt het te verwachten dat de groepen leerlingen die de Eindtoets maken, steeds leerlingen uit groep 8 van het basisonderwijs, van jaar tot jaar niet al te veel in vaardigheid zullen verschillen. Een ander voorbeeld, waarbij we horizontaal willen equivaleren, zijn de eindexamens van het eerste en het tweede tijdvak.

De tweede situatie waarin we zouden willen equivaleren is die waarbij we de prestaties op twee toetsen willen vergelijken die een verschillende moeilijkheidsgraad hebben en dan ook bedoeld zijn voor groepen met verschillende vaardigheidsniveaus. Bij deze zogenaamde verticale equivalering willen we dus corrigeren voor reeds vooraf geplande verschillen in moeilijkheidsgraad tussen de toetsen. Als we bijvoorbeeld

Mavo-C en Mavo-D examens willen equivaleren, dan hebben we te maken met verticale equivalering. Immers, het Mavo-D examen is getracht moeilijker te maken dan het Mavo-C examen terwijl ook de populaties leerlingen in vaardigheid zullen verschillen.

Gezien de extra complicaties (ongelijke moeilijkheid en vaardigheden) zal het duidelijk zijn dat verticaal equivaleren in het algemeen problematischer zal verlopen dan horizontaal equivaleren. Historisch gezien is de theorie van het equivaleren dan ook ontwikkeld voor de situatie waarin we horizontaal willen equivaleren; verticaal equivaleren is pas later ontstaan. Alhoewel er ook binnen het kader van de KTT al enige aandacht aan wordt besteed, is toepassing van verticaal equivaleren eigenlijk pas goed mogelijk als we met IRT werken. We komen hier later nog op terug. In paragraaf 8.1.1 geven we een beknopt overzicht van de psychometrische voorwaarden die in de loop der tijd aan equivalering zijn gesteld. We willen hier reeds opmerken dat in de praktijk niet strikt aan deze voorwaarden wordt vastgehouden. Voor de volledigheid en voor een beter begrip van het equivaleerprobleem worden ze hier toch besproken. Vervolgens bespreken we in paragraaf 8.1.2 de eerste stap van elk equivaleerprobleem: volgens welk design moeten de gegevens die nodig zijn voor het equivaleren, verzameld worden?

8.1.1 Psychometrische voorwaarden voor equivaleren

We kunnen equivaleren als een psychometrisch, maar ook als een statistisch probleem opvatten. We zullen uitleggen wat we hiermee bedoelen. Laten we eerst maar eens aannemen dat we aan een statisticus zonder kennis van de psychometrie vragen om twee toetsen te equivaleren. Daar deze statisticus geen notie van het begrip ware score heeft, is voor hem alleen maar de geobserveerde score van belang. Equivaleren betekent voor hem het zoeken van een relatie tussen de geobserveerde scores van de twee toetsen. Om deze relatie te vinden zal hij bepaalde statistische aannames moeten maken, zoals bijvoorbeeld de aanname dat de geobserveerde scores normaal verdeeld zijn. Vervolgens gebruikt hij een of andere statistische methode om de functionele dan wel structurele relatie tussen de geobserveerde scores vast te leggen. Hoe dit alles precies in zijn werk gaat, is hier niet van belang. De gevolgde werkwijze van de statisticus zullen we statistisch equivaleren noemen. Het moge duidelijk zijn dat equivaleren op deze manier een relatief eenvoudige empirische procedure geworden is: alleen de data en de statistiek zijn hier van belang. De psychometrie wordt in het geheel niet gebruikt. Statistisch equivaleren zoals hierboven beschreven, legt geen

enkele psychometrische restrictie aan de toetsen op. De twee toetsen zouden bijvoorbeeld verschillende betrouwbaarheden kunnen hebben of zelfs verschillende vaardigheden kunnen meten. Als we spreken over (psychometrisch) equivaleren, zullen we dus altijd de psychometrie op de een of andere manier in het verhaal moeten betrekken. Het zal dan ook blijken dat het noodzakelijk is om psychometrische voorwaarden op te leggen aan de te equivaleren toetsen. Bovendien zal blijken dat ook de equivaleerfunctie aan bepaalde voorwaarden moet voldoen. De rest van deze paragraaf zal een beschrijving van deze voorwaarden geven.

Voordat we echter een beschrijving van deze eisen geven, willen we eerst een opmerking maken. Bij het equivaleren van twee toetsen is het, zoals later zal blijken, van groot belang om de betrokken populatie(s) goed te definiëren. De belangrijkste reden hiervoor is dat ook de gebruikte meetmodellen, de KTT en de IRT, altijd met een (of meer) populaties werken. Zo is bijvoorbeeld de betrouwbaarheid van een toets in de KTT populatie-afhankelijk. We komen hier later nog op terug.

Theoretische overwegingen (Angoff, 1971) leiden tot de volgende, vrij algemeen aanvaarde, vier voorwaarden of eisen (Petersen e.a., 1989) met betrekking tot het equivaleren van twee toetsen:

(1) De toetsen moeten dezelfde vaardigheid meten.
(2) De geëquivaleerde scores op de twee toetsen moeten uitwisselbaar zijn.
(3) De equivaleerfunctie moet invariant over groepen personen zijn.
(4) De equivalering moet symmetrisch zijn.

We zullen aangeven wat deze theoretische eisen voor de praktijk van het equivaleren betekenen.

De eerste voorwaarde kan gezien worden als een gezond verstand voorwaarde. Hierbij kunnen we opmerken dat het geen enkele zin heeft om een toets engels met een toets natuurkunde te equivaleren. Dit zou namelijk kunnen leiden tot uitspraken zoals Piets vaardigheid in engels is even groot als Jans natuurkunde vaardigheid. Bij equivaleren met behulp van de KTT zijn er verschillende mogelijkheden om aan de eerste voorwaarde te voldoen. De zwakst mogelijke is die waarbij we eisen dat de twee toetsen congeneriek zijn; de sterkste is die van parallelliteit. Voor meer informatie omtrent de begrippen congeneriek en parallelliteit verwijzen we naar paragraaf 3.6.1 (zie ook tabel 3.1). Op dit moment volstaat de opmerking dat naarmate de voorwaarden die we stellen aan de te equivaleren toetsen strenger worden, de equivalering van de toetsen eenvoudiger en beter wordt. Immers, als de eisen die we stellen om over dezelfde vaardigheid te kunnen spreken sterker worden, gaan de toetsen meer op elkaar lijken: de toetsen zelf worden dan al meer 'equivalent'. Bij equivaleren met behulp van de IRT dient de eerste eis, strikt genomen, vervangen te worden door de

sterkere eis van unidimensionaliteit. We verwijzen voor de betekenis hiervan naar paragraaf 4.3.1. De laatste jaren zijn er echter ook voor meerdimensionale IRT-modellen equivaleermethoden ontwikkeld. Daar deze methoden nooit aan de unidimensionaliteitseis kunnen voldoen, zullen we deze 'quasi-equivalering' noemen. Een voorbeeld hiervan zullen we bespreken in paragraaf 8.3.4.

De tweede voorwaarde, de uitwisselbaarheid van de scores, ook wel de rechtvaardigheidseis genoemd, is oorspronkelijk geformuleerd door Angoff (1971), die er de volgende inhoud aan gaf. Het mag voor personen niet uitmaken welke van de twee geëquivaleerde scores gebruikt worden, bijvoorbeeld om een zak/slaag beslissing te nemen. Angoff werkte in het kader van de KTT en stelde vast dat deze voorwaarde noodzakelijkerwijs parallelliteit van de toetsen veronderstelt. Angoff neemt dus daarmee ook de sterkst mogelijke versie van de eerste eis aan. Maar dat zou betekenen dat we alleen maar parallelle toetsen kunnen equivaleren. Daarom is deze strikte voorwaarde door hem afgezwakt tot het even betrouwbaar zijn van de toetsen.

Lord (1980) heeft de rechtvaardigheidseis voor equivalering met behulp van de IRT gepreciseerd als: twee toetsen X en Y zijn uitwisselbaar of sterk equivalent als geen enkele persoon, met een gegeven vaardigheid, een reden heeft om de ene boven de andere toets te prefereren. Het moge duidelijk zijn dat sterk equivalente toetsen het ideaal is. Dat de constructie van sterk equivalente toetsen echter veelal onmogelijk zal zijn kunnen we eenvoudig aantonen. Beschouw daartoe twee toetsen die elk slechts één item bevatten. Willen deze toetsen sterk equivalent zijn, dan moet voor elke willekeurig gekozen persoon de kans op een goed antwoord voor beide items precies gelijk zijn. Maar dit betekent dat de beide items dezelfde itemparameters moeten hebben, ze moeten dus even moeilijk zijn. In het algemeen zal dus gelden dat twee willekeurige toetsen dan en slechts dan sterk equivalent zijn, als er voor elk item uit de ene toets een item uit de andere toets te vinden is dat gelijke itemparameters heeft en omgekeerd. We zien dan gelijk dat een noodzakelijke voorwaarde hiervoor is dat de toetsen ook precies even lang moeten zijn. De eigenschap dat er voor elk item uit de ene toets een 'vergelijkbaar' item uit de andere toets gevonden kan worden, is wat Samejima (1977) het sterk parallel zijn van twee toetsen noemt. Uit de praktijk blijkt dat het vrijwel onmogelijk is om sterk parallelle (equivalente) toetsen te construeren. Deze observatie heeft dan ook geleid (Divgi, 1981 en Yen, 1983) tot een afzwakking van Lords rechtvaardigheidseis tot: twee toetsen zijn zwak geëquivaleerd als elke persoon in de populatie dezelfde verwachte score op beide toetsen heeft. Merk op dat de gebruikte begrippen sterk en zwak logische benamingen zijn. Uit de definities volgt immers eenvoudig dat sterk geëquivaleerde (sterk parallelle) toetsen ook zwak geëquivaleerd zijn. De bovenstaande overwegingen zijn strikt genomen alleen voor het

equivaleren met behulp van de IRT geldig. Omdat de KTT gezien kan worden als een speciaal geval van de IRT (Lord, 1980), wordt er vaak beweerd dat parallelliteit, maar dan in de KTT betekenis, ook voor equivaleren met behulp van de KTT moet gelden. Maar de KTT houdt zich niet bezig met items, doch met toetsscores zodat het voorgaande zeer de vraag is. Bovendien is het zo dat als we de rechtvaardigheidseis zo strikt zouden nemen als Lord, we voor wat de KTT betreft weer terug zijn bij de aanvankelijke voorwaarde van parallelliteit van Angoff. In de praktijk van het equivaleren zal zowel in de KTT als in de IRT zelden voldaan zijn aan de sterkst mogelijke variant van de uitwisselbaarheidsvoorwaarde; in het algemeen zal slechts aan de besproken zwakke varianten zijn voldaan.

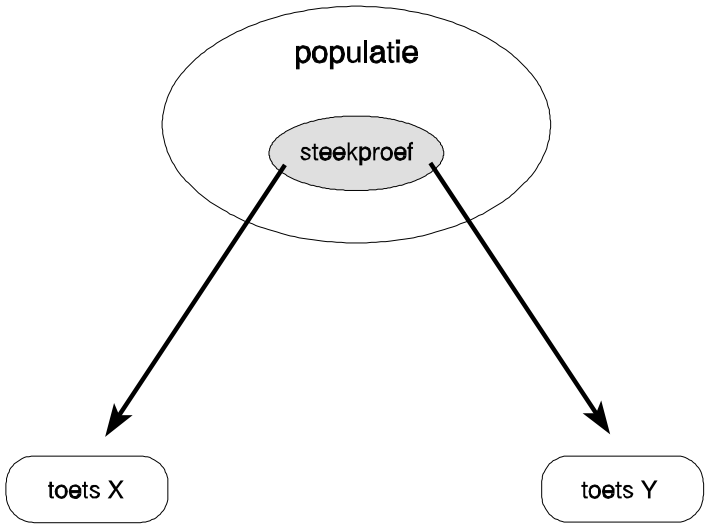
De laatste twee eisen, de invariantie- en symmetrie-eis, zijn het logisch gevolg van het eigenlijke doel van het equivaleren, namelijk het vinden van gelijkwaardige scores. Als scores op twee toetsen gelijkwaardig zijn, dan moet er een één-één relatie bestaan tussen die scores. Maar een één-één relatie is zowel uniek als inverteerbaar. De uniciteit vindt zijn weer-spiegeling in de derde eis, de invariantie over groepen. Als voorbeeld van twee groepen nemen we de opsplitsing van de populatie op basis van sexe. De invariantie eis stelt dan dat de equivaleerfunctie voor de jongens gelijk moet zijn aan die van de meisjes. Als dit niet zo zou zijn, dan is er een score op de ene toets die voor een jongen een andere equivalente score heeft op de tweede toets dan voor een meisje. De twee verschillende equivaleerfuncties hebben één score omgezet in twee verschillende scores. De vierde eis, de symmetrie-eis, kan gezien worden als de inverteerbaarheidsconditie. Stel dat voor een willekeurige score x_0 op toets X een equivalente score y_0 op toets Y gevonden is. De symmetrie eis zegt nu dat als we voor y_0 een equivalente score op toets X zoeken, dat deze score x_0 moet zijn. De derde eis, de invariantie-eis, maakt wederom duidelijk dat we de populatie precies moeten definiëren. Als we namelijk de populatie in het voorgaande definiëren als 'de meisjes', dan is er wat betreft de derde eis wellicht geen probleem meer. We schrijven hier wellicht omdat ook deze populatie weer opgedeeld kan worden, bijvoorbeeld naar leeftijd. Voor de praktijk van het equivaleren betekent dit, dat men er in ieder geval zeker van moet zijn dat de toetsen, in de eventueel te onderscheiden subpopulaties, geen verschillende vaardigheden moeten meten. Dit onderwerp, onzuiverheid, wordt in hoofdstuk 9 besproken. Aan de vierde eis, de symmetrie-eis, kan in de praktijk bijna altijd voldaan worden.

8.1.2 Designs voor equivaleren

De eerste stap die bij equivalering genomen moet worden is het vaststellen van het design voor de verzameling van de data. Voor elk design geldt dat we bij equivaleren altijd uitgaan van een of meer populaties, waaruit een steekproef (of steekproeven) van leerlingen de te equivaleren toetsen maken. Alhoewel we in sommige equivaleerproblemen vrij zijn in de keuze van een design, zij vooraf opgemerkt dat de keuze in de praktijk vaak voor een groot deel wordt bepaald door praktische randvoorwaarden. Bij equivalering wordt veelal gebruik gemaakt van een van de volgende drie basisdesigns, welke in de figuren 8.1, 8.2 en 8.3 schematisch worden weergegeven, het single group design, het random group design en het ankertoetsdesign.

Single group design

Bij dit design maakt één groep leerlingen alle te equivaleren toetsen. Als we twee toetsen willen equivaleren, zeg toets X en toets Y, dan maakt deze groep leerlingen eerst toets X en daarna toets Y. Als vermoeidheidsaspecten een rol spelen, dan is het mogelijk dat toets Y



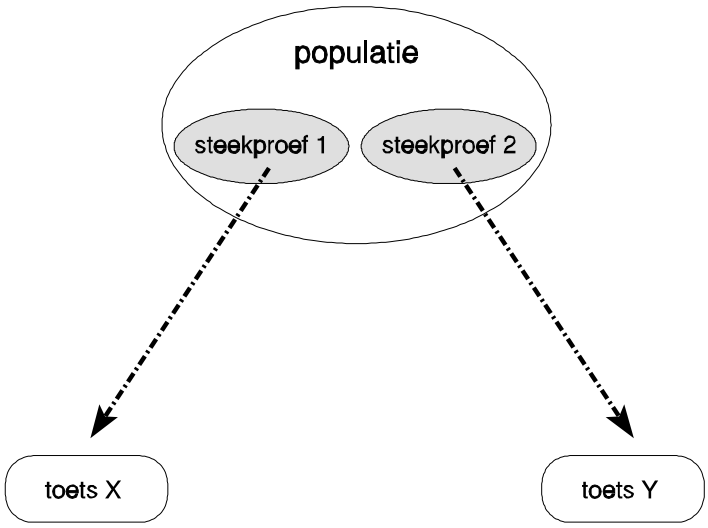
Figuur 8.1
Single group design

relatief moeilijker lijkt dan hij in werkelijkheid is. Anderzijds is het ook mogelijk dat er een zeker oefeneffect optreedt, toets Y lijkt dan gemakkelijker. Om deze effecten te vermijden, wordt bij dit design vaak gebruikt gemaakt van verwisseling: een helft van de kandidaten maakt eerst toets X en daarna Y, terwijl de andere helft eerst toets Y

en daarna X maakt. De idee is uiteraard dat oefen- en vermoeidheidseffecten elkaar dan opheffen. Helaas is het niet goed mogelijk om te onderzoeken of dit inderdaad ook gebeurt. Een ander probleem dat hiermee niet opgelost kan worden is het tijdsduureffect. Als beide toetsen een afnametijd van, zeg drie uur vragen, zal voor de afname van beide toetsen praktisch een hele dag nodig zijn. Bovendien is het vaak zo dat men een nieuwe versie van een toets wil equivaleren met een oudere, zoals bij examens en de Eindtoets. Bij dit design zal de steekproef dus zowel de oude als de nieuwe toets moeten maken. Dit zijn geen gewenste zaken, daarom wordt dit design niet vaak toegepast.

Random group design

Bij dit design, zie figuur 8.2, maken twee aselect getrokken groepen leerlingen uit één populatie elk één toets. De nadelen die we bij het single group design hebben aangegeven, zijn bij het random group design niet aanwezig. Bij nieuwe en oude versies van een toets of examen kan de geheimhouding van de oude echter wel een rol spelen. Bij dit design hebben we de extra aanname gemaakt dat we beschikken over twee vergelijkbare steekproeven, dat



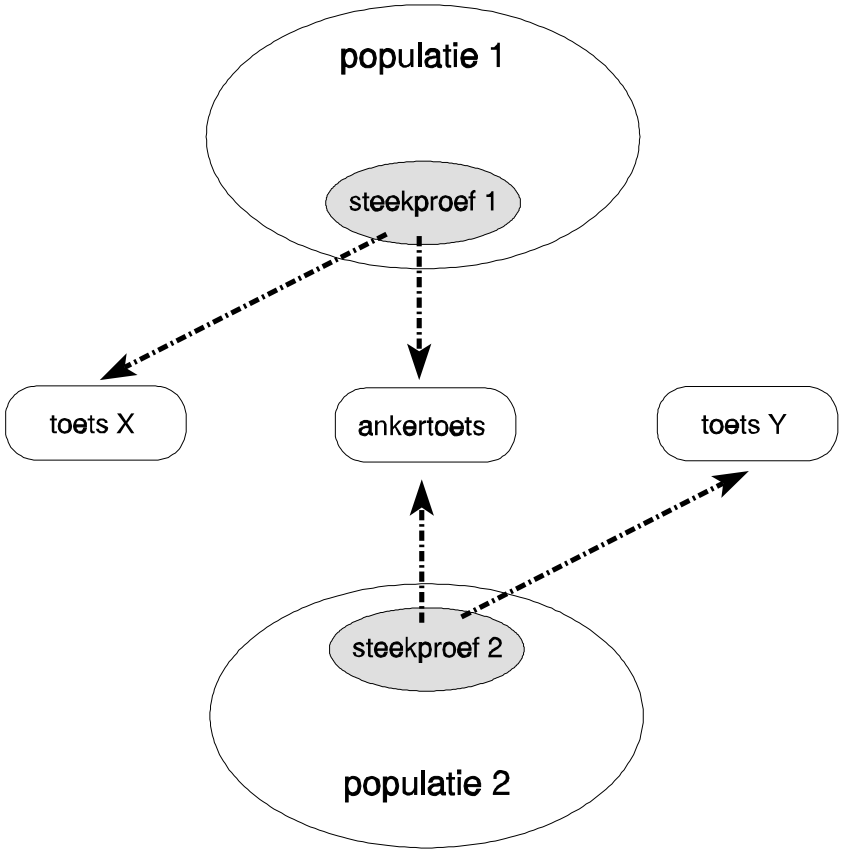
Figuur 8.2
Random group design

wil zeggen met dezelfde vaardigheidsverdeling. Deze vergelijkbaarheid wordt in de praktijk verkregen door slechts één steekproef van leerlingen te trekken en de toetsen daarna aselect toe te wijzen aan de leerlingen. De leerlingen die toets X maken vormen

dan steekproef 1, terwijl steekproef 2 bestaat uit die leerlingen die toets Y maken. Alhoewel op deze wijze de vergelijkbaarheid van de twee steekproeven zeer aanneembaar geworden is, kunnen we deze vergelijkbaarheid niet toetsen.

Ankertoetsdesign

Bij het laatste basisdesign, het ankertoetsdesign, maken twee aselect getrokken groepen leerlingen twee toetsen die een aantal items gemeen hebben. Deze groepen kunnen getrokken zijn uit één populatie, maar ook uit twee verschillende populaties. De variant met twee populaties staat in figuur 8.3. De gemeenschappelijke deeltoets wordt het anker genoemd. De bij de eerste twee basisdesigns genoemde bezwaren zijn bij dit design opgeheven. Immers,
alle leerlingen maken slechts een toets, inclusief de ankertoets. Bovendien biedt de ankertoets de mogelijkheid om voor eventuele verschillen tussen de beide groepen leerlingen te corrigeren. Stel bijvoorbeeld dat de tweede groep gemiddeld hoger scoort op de ankertoets dan de eerste: de tweede groep is dan gemiddeld vaardiger dan de eerste. Deze informatie kunnen we gebruiken om voor een eventueel verschil in moeilijkheidsgraad tussen de toetsen te corrigeren. Hoe dit precies in zijn werk gaat staat beschreven in de volgende paragrafen.



Figuur 8.3
Ankertoetsdesign

Tenslotte willen we een opmerking maken over de status van de ankertoets. Als we gebruik maken van de KTT, zullen we in dit hoofdstuk steeds aannemen dat de ankertoets extern is, dat wil zeggen dat de score op toets X alleen bepaald wordt door de antwoorden op toets X (Y). Het is namelijk ook mogelijk dat de ankertoets opgevat wordt als een deel van de te equivaleren toetsen, hetgeen in de literatuur als intern wordt omschreven. De score op toets X (en ook op Y) bestaat dan dus voor een gedeelte uit het aantal goed gemaakte opgaven uit de ankertoets. Bij het equivaleren in de IRT zullen we steeds veronderstellen dat de ankertoets intern is.

Gezien de bovengenoemde nadelen bij de eerste twee basisdesigns, is het derde basisdesign, het ankertoetsdesign, verreweg het meest gebruikte en bestudeerde equivaleerdesign (Petersen e.a., 1989; Harris & Crouse, 1992). Dit gegeven over de gebruikersfrequentie laat onverlet dat in bepaalde situaties de eerste twee basisdesigns, en dan met name het tweede, best geschikt kunnen zijn. Merk bovendien op dat het tweede en het derde basisdesign voorbeelden zijn van designs die datamatrices geven die onvolledig zijn: elke leerling heeft slechts een gedeelte van de items gemaakt. Zoals reeds in hoofdstuk 6 beschreven is, dienen dit soort designs aan bepaalde voorwaarden

te voldoen om naderhand zinvolle conclusies te kunnen trekken. We komen hier later nog op terug.

Op de drie basisdesigns zijn zeer veel varianten en combinaties ontwikkeld. Zonder volledigheid na te streven noemen we er hier een paar. Het design waarin twee aselect getrokken groepen beide toetsen maken en het design waarbij twee groepen ieder een toets maken terwijl een derde groep beide toetsen maakt, het ankergroepdesign, zijn beide voorbeelden van een combinatie van de basisdesigns. Als variant op het ankertoetsdesign kan ook het, eventueel geblokte, kettingdesign (zie hoofdstuk 6) worden genoemd. Voor alle genoemde designs geldt, zoals we later zullen zien, dat ze voor sommige equivaleermethoden wel en voor andere niet bruikbaar zijn.

Equivaleerdesign van de Eindtoets

We eindigen deze paragraaf met een voorbeeld van een design uit de praktijk. Dit betreft het design van de Eindtoets voor de jaren 1990-1993, welke in figuur 8.4 schematisch is weergegeven. Horizontaal in de figuur staan verschillende ankertoetsen en eindtoetsen (EB met jaar), verticaal de jaargroepen leerlingen. In de figuur is met grijs aangegeven wie welke toetsen maakt.

	anker K	anker L	anker M	anker N	EB90	EB91	EB92	EB93
1990	■				■			
		■			■			
					■			
1991	■					■		
		■				■		
			■			■		
						■		
1992		■					■	
			■				■	
				■			■	
							■	
1993		■						■
			■					■
				■				■
								■

Figuur 8.4
Afnamedesign Eindtoets Basisonderwijs 1990-1993

De Eindtoets wordt ieder jaar bij ongeveer 60% van de leerlingen uit groep 8 van het basisonderwijs afgenomen. Bovendien maakt elk jaar een steekproef van ongeveer 3000 leerlingen, behalve de Eindtoets van hun eigen jaar, een ankertoets. Zo'n ankertoets

is een verkleinde versies (45 items) van de Eindtoets (180 items): zowel qua inhoud alsook qua psychometrische eigenschappen zijn beide toetsen vergelijkbaar. Deze ankertoetsen houden, in tegenstelling tot de Eindtoets, dezelfde samenstelling en dienen louter voor de equivalering. Aangezien de inhoud van de Eindtoets in de loop der tijd aangepast wordt aan het veranderende onderwijs, moeten de ankertoetsen, om nog vergelijkbaar met de Eindtoets te blijven, na verloop van tijd vervangen worden. Voor twee verschillende jaren in welke dezelfde ankertoets afgenomen is, hebben we te maken met een ankertoetsdesign. Het totale design is een voorbeeld van een combinatie van de basisdesigns.

8.2 Equivaleren in de klassieke testtheorie

Voorafgaande aan de bespreking van het equivaleren in de KTT, willen we eerst een algemene opmerking maken omtrent de KTT die in dit verband van belang is. Zoals in hoofdstuk 4 is beschreven, is een van de grootste bezwaren van de KTT de onmogelijkheid om de moeilijkheid van een toets en de vaardigheid van de populatie te scheiden. Of, met andere woorden, alle uit de KTT bekende begrippen zoals *p*-waarden, r_{it} en betrouwbaarheid, hebben steeds betrekking op één populatie en (vaak) één toets. Bij het equivaleren, waar we te maken hebben met verschillende toetsen, eventueel met verschillende moeilijkheden, en (eventueel) met verschillende populaties, kan dit bezwaar ons natuurlijk nog extra parten spelen. Toch wordt equivaleren met behulp van de KTT nog steeds vrij regelmatig gebruikt. Een eerste reden hiervoor is de grote hoeveelheid van beschikbare methoden, die in de praktijk naar tevredenheid van de gebruiker worden toegepast. Een tweede reden is dat in die gevallen waar equivaleren met behulp van de IRT onmogelijk is, men wel met behulp van de KTT moet equivaleren. De eisen die de KTT stelt zijn immers zwakker als die van de IRT.

Binnen de KTT kunnen we grofweg twee klassen van equivaleermethoden onderscheiden. De eerste klasse maakt alleen gebruik van geobserveerde scores (’observed score equating’) terwijl de tweede klasse werkt met ware scores (’true score equating’). In de praktijk worden meestal alleen equivaleermethoden gebruikt die werken met de geobserveerde scores. Een eerste reden hiervoor is de eenvoud. Een tweede, en minstens zo’n belangrijke, reden is dat als men toch wil werken met ware scores, IRT vaak te prefereren is. In het kader van de KTT zullen we ons dan ook zo veel mogelijk beperken tot equivaleermethoden die gebruik maken van geobserveerde scores, soms zullen we echter ook de ware scores in het verhaal betrekken. Hierbij

houden we uiteraard rekening met de psychometrische voorwaarden zoals die in paragraaf 8.1.1 zijn behandeld.

In paragraaf 8.2.1 zullen we de basisequivaleermethoden binnen de KTT beschrijven. Aangaande de voorwaarden uit paragraaf 8.1.1, zullen we altijd aannemen dat we toetsen willen equivaleren die dezelfde vaardigheid meten. Op z'n minst moeten de toetsen dus congeneriek zijn. Een extra psychometrische aanname die vaak gemaakt wordt is dat de toetsen even betrouwbaar zijn. Het belang van gelijke betrouwbaarheid van de toetsen is evident. Zouden de toetsen namelijk niet even betrouwbaar zijn, dan zou een zwakke leerling de voorkeur geven aan een minder betrouwbare toets, terwijl de goede leerling meer baat zou hebben bij de meer betrouwbare toets. Immers, de zwakke leerling heeft bij een slechter meetinstrument een grotere kans om bijvoorbeeld boven de cesuur te scoren. Zelfs aan de zwakke versie van de rechtvaardigheidseis kan dus nooit voldaan worden voor toetsen die niet even betrouwbaar zijn. Bovendien blijkt uit de praktijk dat in de meeste situaties de (geschatte) betrouwbaarheid van de te equivaleren toetsen (ongeveer) gelijk is. Wij zullen de eis van gelijke betrouwbaarheden hier dan ook maken. In paragraaf 8.2.2 zullen we de equivaleermethoden in de KTT voor het ankertoetsdesign bespreken.

8.2.1 Basismethoden voor equivaleren

In de KTT zijn er drie basismethoden in gebruik om een equivaleerfunctie te bepalen tussen twee toetsen: de equipercentiel methode, de lineaire methode en de regressie methode, die we nu achtereenvolgens beschrijven. Om een beter inzicht te krijgen in de problematiek, zullen we in eerste instantie steeds aannemen dat we over de data beschikken van één gehele populatie $\wp$. Daarna zullen we de parameters van de vaardigheidsverdeling in die populatie schatten uit de getrokken steekproef. Een opmerking omtrent de notatie. De te equivaleren toetsen worden aangegeven met hoofdletters (bijvoorbeeld X), terwijl de (geobserveerde) scores op die toetsen cursief genoteerd worden (bijvoorbeeld X).

Equipercentielmethode

De equipercentielmethode werkt als volgt: Kies de equivaleerfunctie zodanig dat de scores op de toetsen X en Y geëquivaleerd zijn als ze corresponderen met dezelfde percentiele rang in de populatie, waaronder we verstaan het percentage scores in de

populatie dat gelijk of kleiner is. Deze equivaleermethode is historisch gezien de belangrijkste en werd vroeger zelfs als definitie gehanteerd: "Two scores, one on form X and the other on form Y (where X and Y measure the same function with the same degree of reliability), may be considered equivalent if their corresponding percentile ranks in any group are equal" (Lord, 1950; Flanagan, 1951).

Laat dus nu $\mathcal{P}$ een populatie van leerlingen zijn waarvoor de te equivaleren toetsen X en Y geschikt zijn. Elke leerling uit $\mathcal{P}$ kan dus getoetst worden met X en/of Y. Stel dat $F(x)$ en $G(y)$ de verdelingsfuncties van de geobserveerde scores van de toetsen X en Y in de populatie $\mathcal{P}$ zijn, dat wil zeggen

$$\begin{cases} F(x) = \text{proportie leerlingen in } \mathcal{P} \text{ met } X \leq x, \\ \\ G(y) = \text{proportie leerlingen in } \mathcal{P} \text{ met } Y \leq y. \end{cases} \tag{8.1}$$

Bij de equipercentielmethode worden alle percentiele rangen gelijkgesteld, hetgeen natuurlijk alleen mogelijk is als voor een willekeurige waarde van een percentiele rang p^* met $p^* = 100p$ geldt dat

$$F(x) = p \quad \text{en} \quad G(y) = p. \tag{8.2}$$

Het is eenvoudig na te gaan dat voor strikt monotone F en G er dan geldt dat

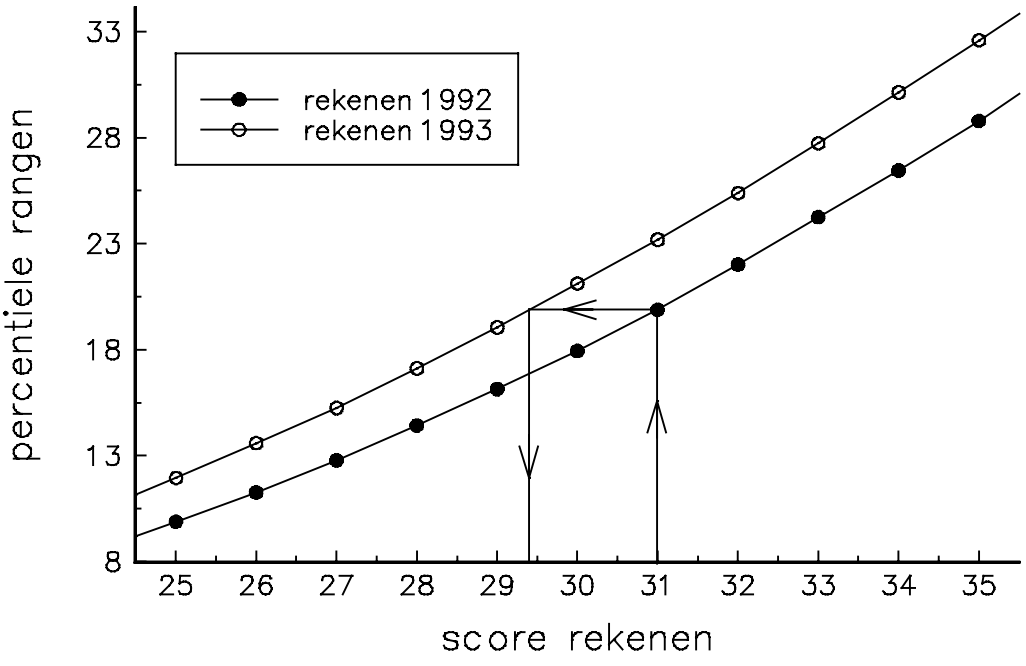
$$x = F^{-1}(G(y)). \tag{8.3}$$

De inverse functie van F, F^{-1}, wordt gegeven door het voorschrift dat $F^{-1}(p)$ die waarde van x is waarvoor geldt dat $F(x) = p$. Merk op dat x nu een functie van y geworden is. We geven dit aan met $e_X(y)$. Dus $e_X(y) = F^{-1}(G(y))$ equivaleert X en Y op $\mathcal{P}$, waarbij $e_X(y)$ de equivaleerfunctie is. Nu kan de percentiele rang p^* elke mogelijke waarde tussen 0 en 100 aannemen. De scores hebben echter slechts een eindig bereik daar alleen scores tussen 0 (alle items fout) en de maximale score (alle items goed) mogelijk zijn. De verdelingsfuncties F en G zijn dan niet meer strikt monotoon. Maar dit betekent ook dat de inverse functie nooit helemaal exact bekend is en dat de waarde van de inverse functie op de onbekende plaatsen op de een of andere manier moet worden ingevuld. Dit proces van invullen staat bekend onder de naam 'smoothen'. We zullen dit later aan de hand van een voorbeeld demonstreren.

Een ander moeilijk probleem blijft natuurlijk het bepalen van $F(x)$ en $G(y)$ omdat we in de praktijk nooit over de gehele populatie $\mathcal{P}$, maar slechts over steekproeven uit $\mathcal{P}$ beschikken. We zullen ons dus altijd moeten behelpen met schattingen van de functies F en G. Bovendien moeten complete verdelingsfuncties met in principe oneindig veel parameters geschat worden. Als we over een aselecte steekproef beschikken, dan zou

als schatting van de verdelingsfunctie natuurlijk de geobserveerde kunnen dienen. De geobserveerde verdelingsfunctie is eenvoudig uit de geobserveerde frequentieverdeling te construeren en kan bovendien met veel statistische pakketten uitgerekend worden. Hoe goed deze schatting is, hangt uiteraard af van de populatie, de steekproef en de toetsen. Het moge duidelijk zijn dat bij grotere steekproeven de geschatte verdelingsfunctie de ware beter zal benaderen.

Als voorbeeld zullen we nu laten zien hoe twee versies van de Eindtoets met behulp van de equipercentielmethode geëquivaleerd kunnen worden. We zullen ons hier beperken tot het onderdeel rekenen (60 items) voor de jaren 1992 en 1993. Als eerste stap moeten we dan de beschikking hebben over één populatie $\wp$. We kunnen dit doen als we de populatie $\wp$ definiëren als 'alle kinderen die in een willekeurig jaar in groep 8 zitten'. In werkelijkheid beschikken we natuurlijk niet over $\wp$, maar slechts over twee steekproeven van leerlingen, een die aan de Eindtoets van 1992 en een die aan de Eindtoets van 1993 deelnam; beide steekproeven bevatten ongeveer 100.000 leerlingen. De verdelingsfunctie van de scores van 1992 noemen we G en die van 1993 noemen we F. Merk op dat bij de Eindtoets de scores gegeven worden door middel van het aantal goed beantwoorde opgaven. Daar we over een zeer grote steekproef beschikken, mogen we aannemen dat de geobserveerde verdelingsfunctie $\hat{G}$ een goede schatting is van G. Hetzelfde verhaal gaat natuurlijk op voor F en $\hat{F}$. De geobserveerde verdelingsfuncties $\hat{F}$ en $\hat{G}$ zijn voor scores tussen 25 en 35 weergegeven in figuur 8.5. We hebben bovendien beide verdelingsfuncties een vloeiend verloop gegeven, dat wil zeggen een nette benaderende lijn door de discrete verdelingsfunctie getrokken. Dit is



Figuur 8.5
Equipercentiel equivaleren Eindtoets

wat we hiervoor smoothen genoemd hebben. Merk op dat de verdelingsfunctie van 1993, voor de gegeven scores, overal boven die van 1992 ligt. Ook voor de niet gepresenteerde scores bleek dit zo te zijn. Uit de aanname dat beide steekproeven getrokken zijn uit een en dezelfde populatie volgt dus dat de Eindtoets van 1993 moeilijker is dan de Eindtoets van 1992, uiteraard voor het onderdeel rekenen. Nu zijn alle gegevens voor de equipercentiel equivalering beschikbaar. Beschouw nu bijvoorbeeld een score van 31 op het onderdeel rekenen van de Eindtoets van 1992; bij deze score hoort een percentiel van (ongeveer) 20. Bij datzelfde percentiel zou een reken score op de Eindtoets van 1993 van (ongeveer) 29.4 horen. Maar deze score kan niet voorkomen, zodat we de dichtstbijzijnde score kiezen, of, met andere woorden, we ronden 29.4 af tot 29.

Lineaire methode

De lineaire methode kan omschreven worden met de volgende regel: 'Kies de equivaleerfunctie zodanig dat twee scores op X en Y equivalent zijn als ze hetzelfde aantal standaarddeviaties afwijken van de gemiddelden, ofwel dezelfde standaardscore hebben'. Voor toets X (Y) duiden we het gemiddelde van de geobserveerde scores in de populatie $\wp$ aan met μ_X (μ_Y) en de standaarddeviatie van de scores met σ_X (σ_Y). Het gelijk stellen van de standaardscores is dan equivalent met

$$\frac{X - \mu_X}{\sigma_X} = \frac{Y - \mu_Y}{\sigma_Y}. \tag{8.4}$$

Herschikken van termen in (8.4) geeft dan direct de formule voor het lineair equivaleren:

$$e_X(Y) = \mu_X + \frac{\sigma_X}{\sigma_Y}(Y - \mu_Y). \tag{8.5}$$

We merken hier op, dat als we de sterke variant van de rechtvaardigheidseis in de KTT, de toetsen zijn parallel, zouden hebben aangenomen, dat (8.4) dan reduceert tot $X = Y$. De scores op de toetsen zijn dan dus per definitie lineair geëquivaleerd.

Lineair equivaleren kan ook gezien worden als een bijzonder geval van equipercentiel equivaleren in de zin dat slechts de eerste twee momenten van de scoreverdelingen gelijkgesteld worden (Braun & Holland, 1982). Er kan namelijk eenvoudig aangetoond

worden dat bij equipercentiel equivaleren alle momenten aan elkaar gelijk gesteld worden. Een extra aanname bij lineaire equivalering is dus dat de hogere momenten van de scoreverdelingen van beide toetsen identiek zijn. Deze benadering start dan ook met de aanname dat F en G schaalinvariante functies zijn. Schaalinvariante functies zijn functies waarvan de ene functie op een lineaire transformatie na, gelijk is aan de andere. Met andere woorden, schaalinvariante functies hebben dezelfde vorm. Bij equipercentiel equivaleren moeten complete verdelingsfuncties geschat worden, hetgeen een groot nadeel van die methode is. Omdat het in het algemeen beter is om minder dan meer parameters te schatten, verdient lineair equivaleren, daar waar toepasbaar, de voorkeur.

Net zoals bij het equipercentiel equivaleren, zijn ook bij het lineair equivaleren de populatie gegevens, in dit geval de gemiddelden en de standaarddeviaties, niet bekend. Deze moeten dus altijd uit de data geschat worden en vervolgens ingevuld worden in (8.5). Als schatters voor μ_X en σ_X komen bijvoorbeeld de steekproefmomenten $\bar{X}$ en s_X in aanmerking.

Als de toetsen X en Y niet even betrouwbaar zijn, kunnen we ook lineair equivaleren. Het is duidelijk dat we nu niet meer alleen met geobserveerde scores uit de voeten kunnen. De betrouwbaarheid is immers een functie van zowel de ware als van de geobserveerde scores. De ware scores dienen nu dus op de een of andere manier expliciet gebruikt te worden. De simpelste manier is nu om (8.4) te herschrijven tot een vergelijking tussen de ware scores. Hiertoe dienen we dan zowel de geobserveerde scores als ook de parameters van de geobserveerde variabelen te vervangen door de ware score equivalenten. Dus voor toets X vervangen we μ_X door $\mu_{T(X)}$ en σ_X door $\sigma_{T(X)}$; voor toets Y geldt hetzelfde. Dit levert dan

$$\frac{T(\mathrm{X}) - \mu_{T(X)}}{\sigma_{T(X)}} = \frac{T(Y) - \mu_{T(Y)}}{\sigma_{T(Y)}}. \tag{8.6}$$

Merk nu op dat alle termen in (8.6) onbekend zijn. Zowel de ware scores $T(\mathrm{X})$ en $T(\mathrm{Y})$ als ook de parameters $\mu_{T(X)}$, $\sigma_{T(X)}$, $\mu_{T(Y)}$ en $\sigma_{T(Y)}$ van de ware score verdelingen zijn niet bekend. Gelukkig beschikken we voor alle onbekenden over goede schatters. Voor het gemak beperken we ons in de schrijfwijze even tot toets X. We starten met de parameters, daar deze het eenvoudigst zijn. Immers, uit hoofdstuk 3 weten we dat $\mu_{T(X)} = \mathscr{E}(T) = \mathscr{E}(X) = \mu_X$ en $\sigma^2_{T(X)} = \sigma^2_X \rho_{XX'}$. Voor de schatting van de ware scores beschikken we over twee kandidaten: de geobserveerde-score-schatter en de Kelley-schatter. Als we de geobserveerde score nemen als schatter voor

de ware scores, dan vullen we voor $T(X)$ dus X in. Invullen van deze schattingen in (8.6) levert dan

$$\frac{X - \mu_X}{\sqrt{\rho_{XX'}}\,\sigma_X} = \frac{Y - \mu_Y}{\sqrt{\rho_{YY'}}\,\sigma_Y}. \tag{8.7}$$

Herschikking van de termen in (8.7) levert dan de eerste formule voor het lineair equivaleren van twee niet even betrouwbare toetsen:

$$e_X(Y) = \mu_X + \frac{\sigma_X\sqrt{\rho_{XX'}}}{\sigma_Y\sqrt{\rho_{YY'}}}(Y - \mu_Y). \tag{8.8}$$

Als we de Kelley-schatter nemen als schatter van de ware score, dan wordt de schatter van de teller van het linkerlid van (8.6) gegeven door

$$\frac{\sigma^2_{E(X)}}{\sigma^2_{E(X)} + \sigma^2_{T(X)}}\,\mu_{T(X)} + \frac{\sigma^2_{T(X)}}{\sigma^2_{E(X)} + \sigma^2_{T(X)}}\,X - \mu_{T(X)}, \tag{8.9}$$

waarbij $\sigma^2_{E(X)}$ de foutenvariantie weergeeft. Uitwerken van (8.9) geeft $\rho_{XX'}(X - \mu_{T(X)})$. Invullen hiervan en van de bovengenoemde schatters voor de parameters en herschikking van de verschillende termen levert dan de tweede formule voor het equivaleren van twee niet even betrouwbare toetsen:

$$e_X(Y) = \mu_X + \frac{\sigma_X\sqrt{\rho_{YY'}}}{\sigma_Y\sqrt{\rho_{XX'}}}(Y - \mu_Y). \tag{8.10}$$

Merk op dat in de formules (8.8) en (8.10) de ratio tussen de wortels van de beide betrouwbaarheden is omgekeerd. Bovendien geldt voor beide formules dat het verschil met (8.5) alleen zit in de ratio van die wortels. Hieruit lezen we dan ook direct af dat het voor twee bijna even betrouwbare toetsen, het praktisch geen verschil maakt of formule (8.5) dan wel (8.8) of (8.10) gebruikt wordt. Ten overvloede wellicht, zullen in de praktijk zowel in (8.8) als in (8.10) schattingen voor de parameters moeten worden ingevuld. Merk op dat nu ook de verschillende betrouwbaarheden geschat moeten worden. Hoe de betrouwbaarheid van een toets geschat kan worden is reeds uitgebreid behandeld in paragraaf 3.6, we zullen dit hier dan ook niet herhalen.

Regressiemethode

Bij de regressiemethode wordt de equivaleerfunctie tussen de scores bepaald door de regressie van de scores van de ene toets op de andere te bepalen. Voor de lineaire regressie van X op Y volgt dan

$$e_X(Y) = \mu_X + \rho_{XY}\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y), \tag{8.11}$$

waarbij ρ_{XY} de correlatie tussen de scores van de toetsen X en Y is. Merk op dat (8.11) identiek is aan (8.8) op de factor ρ_{XY} na. Om ρ_{XY} te schatten is het noodzakelijk om over een steekproef van leerlingen te beschikken die zowel toets X als toets Y gemaakt hebben. Dit is bijvoorbeeld mogelijk als de data verzameld zijn volgens het eerste basisdesign, het single group design. In (8.11) wordt de equivaleerfunctie bepaald door de regressie van X op Y. Als we de rol van X en Y omdraaien, dat wil zeggen als we de regressie van Y op X bepalen, dan vinden we

$$e_Y(X) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X - \mu_X). \tag{8.12}$$

Nu is (8.12) niet de inverse van (8.11), hetgeen niet strookt met de symmetrie eis. De equipercentiel en de lineaire methode voldoen wel aan de symmetrie eis, hetgeen direct uit (8.2) en (8.4) kan worden afgelezen. De regressiemethode dient dus altijd met de nodige voorzichtigheid betracht te worden.

We vervolgen nu het voorbeeld van de equivalering van de Eindtoets. Voor de lineaire equivalering hebben we alleen maar de eerste twee momenten nodig. Schattingen van μ_X etcetera worden uiteraard gegeven door de steekproefmomenten, deze zijn $\bar{X} = 41.22$, $s_X = 11.46$, $\bar{Y} = 41.96$ en $s_Y = 10.98$. Invullen van deze schattingen in (8.5) levert dan de equivaleerfunctie $e_X(Y) = 41.22 + 1.04(Y - 41.96) = -2.57 + 1.04\,Y$. Merk op dat Y bij 1992 hoort en X bij 1993. Voor de score van 31 op de Eindtoets van 1992 vinden we dan de lineair geëquivaleerde score van 29.67 op de Eindtoets van 1993, hetgeen redelijk overeenkomt met de score van 29.4 bij het equipercentiel equivaleren. Merk op dat er voor de regressiemethode nooit genoeg gegevens zijn. Er zijn immers geen leerlingen die beide versies van de Eindtoets gemaakt hebben, zodat we ρ_{XY} niet kunnen schatten.

In het bovenstaande hebben we de equivalering van de toetsen X en Y steeds eerst beschouwd op de totale populatie $\mathcal{P}$. We merkten daarbij op dat we in werkelijkheid nooit beschikken over de gehele populatie, doch slechts uit steekproeven hieruit. We moeten dus altijd de data, en daarmee het design meenemen om tot een goede keuze voor de equivaleerprocedure te komen. Bovendien kan er sprake zijn, zoals bijvoorbeeld bij het verticaal equivaleren, van meerdere populaties. Vooral dit laatste is nog

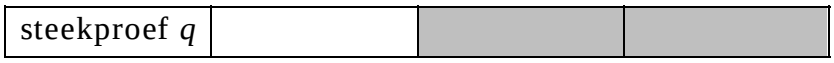
een behoorlijk probleem. Bij de bespreking van het voorbeeld van de equivalering van de Eindtoets hebben we dit probleem een beetje verdoezeld. We hadden daar immers ook twee populaties, die van 1992 en die van 1993, die we samengevoegd hebben tot een (alle leerlingen in groep 8). Dit samenvoegen tot een populatie is statistisch goed gefundeerd (Braun & Holland, 1982), maar conceptueel moeilijk voorstelbaar. Deze populatie heet in de literatuur 'synthetic population'. We zullen in het vervolg dan ook aannemen dat, indien er twee populaties in het geding zijn, deze samengevoegd zijn tot één synthetische populatie. We bespreken nu de equivalering van het in de praktijk vaak voorkomende ankertoetsdesign.

8.2.2 Equivaleren met behulp van het ankertoetsdesign

In deze paragraaf bespreken we het equivaleren indien de data verzameld zijn met een anker- toetsdesign. De nadruk zal hierbij liggen op de meest gebruikte vorm van equivaleren, namelijk lineair equivaleren. Voor de duidelijkheid hebben we het ankertoetsdesign nogmaals weergeven in figuur 8.6. Steekproef *p*, uit populatie 1, maakt toets X en de ankertoets A (X-groep), terwijl steekproef *q* uit populatie 2, toets Y en ankertoets A maakt (Y-groep). De totale steekproef, *p* en *q* samen, zullen we *t* noemen. Populatie 1 en populatie 2 vormen samen de synthetische populatie $\wp$; *t* is een steekproef uit $\wp$.

Allereerst een opmerking over de ankertoets A. Evenals voor de te equivaleren toetsen X en Y, zullen we ook aan de ankertoets psychometrische eisen moeten opleggen. Als we bijvoorbeeld twee toetsen engels willen equivaleren, dan mogen we van de ankertoets op z'n minst verwachten dat deze ook engels meet. Een redelijke eis is dan hier ook dat de ankertoets A congeneriek is met X (en dus ook met Y). Ook hier geldt weer, dat naarmate de eisen sterker worden, de equivalering eenvoudiger wordt. Een overzicht van alle mogelijke psychometrische eisen voor lineair equivaleren die in een ankertoetsdesign gesteld kunnen worden is te vinden in MacCann (1990). Bedenk bovendien dat we steeds veronderstellen dat de ankertoets extern is, zodat de ankertoets niets aan de te equivaleren scores bijdraagt. Uiteraard nemen we weer aan dat toets X en toets Y even betrouwbaar zijn.

	toets X	ankertoets A	toets Y
steekproef *p*			



Figuur 8.6
Ankertoetsdesign

We gaan nu verder met het beschrijven van de equivalering in het ankertoetsdesign. Een belangrijke observatie is nu dat we direct zouden kunnen equivaleren als we over data zouden beschikken in de lege cellen. We zijn dan immers weer terug in de situatie van volledige data uit de vorige paragraaf. Alle equivaleermethoden welke met ontbrekende data werken, vullen dan ook op de een of andere manier deze ontbrekende data in, om zo weer in het volledige data geval terecht te komen. De idee bij dit invullen is natuurlijk om de gegevens van ankertoets A te gebruiken om de scores van leerlingen uit de Y-groep (X-groep) op toets X (Y) te voorspellen. Soms hebben we echter niet de scores op de toetsen nodig, maar kunnen we met minder gegevens volstaan. Als we, bijvoorbeeld, lineair willen equivaleren, dan leert inspectie van (8.5) dat de enige relevante grootheden de gemiddelden en de standaarddeviaties van de scores in de verschillende populaties zijn. Het bepalen van deze gemiddelden en standaarddeviaties, of meer algemeen voor de ingevulde waarden, gebeurt dan uiteraard onder bepaalde aannames. De meest gebruikte aanname is die welke in de literatuur 'constancy of regression' wordt genoemd. Bij deze aanname wordt eerst verondersteld dat de scores op de toetsen X en Y een lineair verband hebben met de ankertoets, zodat lineaire regressie zinvol wordt. Vervolgens veronderstelt men dat de intercept, de regressiecoëfficiënt en de variantie van de schattingsfout van de scores op toets X (Y) op A is gelijk voor de X-groep (Y-groep) en de totale groep (= X-groep + Y-groep). Met andere woorden, als we de totale steekproef zouden hebben geobserveerd, dan zouden we dezelfde schattingen voor alle regressie-parameters gevonden hebben als we nu voor de gedeeltelijke steekproef gevonden hebben.

We zullen nu laten zien waarop de 'constancy of regression' aanname gebaseerd is. Laat daartoe μ_X en σ_X het onbekende gemiddelde en de standaarddeviatie van de scores van toets X zijn in de synthetische populatie Ω. We zullen eerst laten zien hoe we op een eenvoudige manier een goede schatter van μ_X kunnen construeren. Een eerste schatting is simpel te maken. Kies daartoe gewoon het gemiddelde van X in de geobserveerde steekproef p, oftewel $\hat{\mu} = \bar{X}_p$. Het moge duidelijk zijn dat we om deze schatting te kunnen verbeteren op de een of andere manier gebruik zullen moeten maken van de gegevens omtrent A in de totale steekproef t. Daartoe beschouwen we eerst de volledige data (X,A) in steekproef p, waarbij we aannemen dat er een lineair verband is tussen X en A. Stel nu eens dat $X_v = \beta_0 + \beta_1 A_v + \varepsilon_v$ met $\varepsilon_v \sim N(0, \sigma^2)$ in steekproef p. Hierbij, en in het vervolg, staat de subscript v voor een leerling. De

subscripten X, Y, A, p, q en t spreken voor zich; ze verwijzen naar de toetsen en de steekproeven (of bijbehorende populaties). Dan worden de kleinste kwadraten schatters $\hat{\beta}_0$ en $\hat{\beta}_1$ gegeven door

$$\hat{\beta}_0 \equiv b_0 = \bar{X}_p - b_1 \bar{A}_p \quad \text{en}$$
$$\hat{\beta}_1 \equiv b_X = r_{X A_p} s_{X_p} / s_{A_p}, \tag{8.13}$$

waarbij $r_{X A_p}$ de correlatie tussen X en A in steekproef p is. De geschatte waarde van X_v in steekproef p wordt dan, met de gebruikelijke notatie voor gemiddelden, gegeven door

$$\hat{X}_v = \bar{X}_p + b_X (A_v - \bar{A}_p). \tag{8.14}$$

Vervolgens nemen we aan dat deze formule ook geldt voor leerlingen in steekproef q. Met behulp van bovenstaande regressievergelijking kunnen we dus ook voor leerlingen in steekproef q geschatte waarden voor X_v berekenen (imputeren). Merk op dat dit volledig identiek is aan het voorspellen van de waarde van de afhankelijke variabele voor een nieuwe waarde van de onafhankelijke variabele in een eenvoudig regressieprobleem.

Het geschatte gemiddelde in de totale steekproef t wordt gegeven door formule (8.14) te middelen over de totale steekproef t, zodat we vinden

$$\hat{\mu}_X = \bar{X}_p + b_X (\bar{A}_t - \bar{A}_p). \tag{8.15}$$

Dit nieuwe geschatte gemiddelde $\hat{\mu}_X$ is dus verkregen door de gegevens van de steekproeven p en q op een eenvoudige manier samen te nemen. Op dezelfde manier, maar met meer schrijfwerk wat we hier achterwege zullen laten, kunnen we ook een schatting voor σ_X^2 construeren:

$$\hat{\sigma}_X^2 = S_{X_p}^2 + b_X^2 (S_{A_t}^2 - S_{A_p}^2). \tag{8.16}$$

Dit extra schrijfwerk is een rechtstreeks gevolg van het feit dat de standaardfout voor de geïmputeerde waarden anders (en groter) is dan voor de geobserveerde waarden. Op precies dezelfde manier als voor toets X kunnen we ook het (geschatte) gemiddelde en de standaarddeviatie voor toets Y in de totale steekproef t berekenen. Deze worden dan gegeven door

$$\hat{\mu}_Y = \bar{Y}_q + b_Y (\bar{A}_t - \bar{A}_q) \quad \text{en}$$

$$\hat{\sigma}_Y^2 = S_{Y_q}^2 + b_Y^2(S_{A_t}^2 - S_{A_q}^2), \tag{8.17}$$

waarbij b_Y de (geschatte) regressiecoëfficiënt is van Y op A in steekproef q.
Bekijk nu nogmaals de 'constancy of regression' aanname. Als we deze aanname voor toets X in formule vorm opschrijven, dan vinden we

$$\begin{aligned} &\mu_{X_t} - \beta_{XA_t}\mu_{A_t} = \mu_{X_p} - \beta_{XA_p}\mu_{A_p} && \text{intercept} \\ &\beta_{XA_t} = \beta_{XA_p} && \text{regressie-coëfficiënt} \\ &\sigma_{X_t}^2(1 - r_{XA_t}^2) = \sigma_{X_p}^2(1 - r_{XA_p}^2) && \text{foutenvariantie.} \end{aligned} \tag{8.18}$$

Hierbij staan aan de linkerkant steeds de parameters voor de totale steekproef t en aan de rechterkant voor steekproef p. Substitutie van de tweede vergelijking van (8.18) in de eerste en herschikking van de termen levert dan

$$\mu_{X_t} = \mu_{X_p} + \beta_{XA_p}(\mu_{A_t} - \mu_{A_p}). \tag{8.19}$$

Als we wederom in (8.18) de tweede vergelijking in de derde invullen, en bedenken dat $r_{XA_t} = \beta_{XA_t}\sigma_{X_p}/\sigma_{X_t}$, levert herschikken

$$\sigma_{X_t}^2 = \sigma_{X_p}^2 + \beta_{XA_p}(\sigma_{A_t}^2 - \sigma_{A_p}^2). \tag{8.20}$$

Als we nu in de rechterleden van (8.19) en (8.20) de gebruikelijke schattingen voor de parameters substitueren, dan vinden we weer (8.15) en (8.16) terug.
De 'constancy of regression' aanname is dus niets anders dan datgene wat we in een eenvoudig lineair regressieprobleem doen, als we voor het voorspellen van de afhankelijke variabele, waarden van de predictor invullen die niet gebruikt zijn bij het bepalen van de regressievergelijking.

We beschikken nu over de benodigde gegevens om in t tot de eigenlijke equivalering over te gaan. We hebben nu immers voor elke leerling een score (geobserveerd dan wel geïmputeerd) op zowel toets X als op toets Y; bovendien beschikken we nu over (schattingen) van de gemiddelden en van de standaarddeviaties van de scores. In principe kunnen nu alle klassieke equivaleermethoden direct worden uitgevoerd. Voor lineair equivaleren moeten we de gegevens uit de formules (8.15), (8.16) en (8.17) invullen in formule (8.5). Dit levert dan

$$e_X(Y) = \bar{X}_p + b_X\left(\bar{A}_t - \bar{A}_p\right) + \sqrt{\frac{S_{X_p}^2 + b_X^2\left(S_{A_t}^2 - S_{A_p}^2\right)}{S_{Y_q}^2 + b_Y^2\left(S_{A_t}^2 - S_{A_q}^2\right)}}\left(Y - \bar{Y}_q - b_Y\left(\bar{A}_t - \bar{A}_q\right)\right). \qquad (8.21)$$

Bedenk dat we hiervoor steeds aangenomen hebben dat de toetsen X en Y even betrouwbaar zijn. Ook voor toetsen die niet even betrouwbaar zijn, kunnen we, net zoals in paragraaf 8.2.1, een formule voor het lineair equivaleren in het ankertoetsdesign afleiden. Ook dan geldt weer, dat het voor de praktijk weinig verschil uitmaakt of de toetsen even betrouwbaar, danwel bijna even betrouwbaar zijn (MacCann, 1990). Bovendien hebben we aangenomen dat de toetsen X, Y en A congeneriek zijn. Zoals reeds in hoofdstuk 3 is opgemerkt, dient het toetsen op het congeneriek, of het even betrouwbaar, zijn van twee toetsen in een ruimer model plaats te vinden, bijvoorbeeld in een LISREL kader (Jöreskog & Sörbom, 1989). Hiervoor is het echter noodzakelijk om over de covariantie- of correlatiematrix van de toetsscores te beschikken. Omdat in het ankertoetsdesign de toetsen X en Y nooit bij dezelfde leerlingen zijn afgenomen, kunnen we de correlatie tussen X en Y niet schatten. Alleen door extra dataverzameling kunnen we op het congeneriek of even betrouwbaar zijn toetsen. We zullen hier verder niet op ingaan.

We sluiten nu het voorbeeld van de equivalering van de Eindtoets, voor het onderdeel rekenen, af. Daar we over drie verschillende ankertoetsen beschikken, (L, M en N) kunnen we ook op drie verschillende manieren equivaleren. We kunnen namelijk elke ankertoets de rol van A laten spelen in formule (8.21). We zullen de gegevens presenteren voor de ankers L en M. Als we deze formule uitwerken, waarvan we de details hier niet zullen presenteren, dan vinden we voor anker L de equivaleerfunctie $e_X(Y) = 1.04\,Y - 1.82$. Voor anker M wordt de equivaleerfunctie gegeven voor $e_X(Y) = 1.04\,Y - 2.52$. Merk op dat, alhoewel deze formules veel op elkaar lijken, ze toch niet geheel identiek zijn. Het lijkt er dus op dat de invariantie-eis hier geschonden is, daar de equivaleerfuncties voor twee verschillende groepen niet gelijk zijn. Als we echter toetsen of deze twee equivaleerfuncties verschillen, dan blijkt dat ze (statistisch) niet te onderscheiden zijn. Immers, het moge duidelijk zijn dat de standaardfout horende bij (8.21) best behoorlijk groot kan zijn. De equivaleerfunctie is namelijk opgebouwd uit heel veel verschillende elementen, die we allemaal moeten schatten. De fouten die we hierbij maken werken natuurlijk door in het uiteindelijke resultaat. De precieze berekening van de standaardfout van (8.21) is nogal ingewikkeld, en zullen we hier dan ook achterwege laten, zie bijvoorbeeld Braun en Holland (1982). We willen hier nog opmerken dat in de praktijk van de equivalering van de Eindtoets gewerkt wordt met de gemiddelde equivaleerfunctie. Zoals hiervoor al is opgemerkt, hebben we bij de afleiding van (8.21) aangenomen dat de twee te equivaleren toetsen gelijke

betrouwbaarheden hebben. Dit blijkt voor dit voorbeeld redelijk te kloppen. Voor de Eindtoets van 1992 vinden we als schatting van de betrouwbaarheid .918, terwijl we .920 voor die van 1993 vinden, uiteraard steeds voor het onderdeel rekenen.

8.3 Equivaleren met itemresponstheorie

Bij de bespreking van de equivaleermethoden in de KTT hebben we opgemerkt dat het soms problematisch is om de scores van verschillende toetsen op dezelfde schaal uit te drukken, en dus vergelijkbaar te maken, aangezien de moeilijkheid van opgaven of toetsen en de vaardigheid van personen niet gescheiden kunnen worden. In de IRT ligt de zaak heel anders: vaardigheden van personen en kenmerken van items worden middels aparte parameters in een kansmodel aan elkaar gerelateerd. En indien voor een verzameling opgaven in een bepaalde populatie een itemresponsmodel geldt, dan kunnen de vaardigheidsparameters van personen op eenzelfde schaal geschat worden door slechts deelverzamelingen van de betrokken opgaven te beschouwen. Maar dit laatste is nu juist waar het bij de equivalering om gaat. Immers, bij equivalering willen we de scores op verschillende toetsen vergelijkbaar maken. Maar als we de vaardigheidsparameter onafhankelijk van de toetsen kunnen bepalen, hoeven we de scores niet meer vergelijkbaar te maken. Ze liggen immers direct op de vaardigheidsschaal waarop we kunnen weergeven.

Het voorgaande suggereert dat er bij toepassing van de IRT geen equivaleerproblemen zijn. In principe is deze uitspraak juist, maar er zijn in de praktijk nog diverse interessante problemen, die we nu kort aan zullen duiden.

Allereerst moet er voldaan zijn aan de eerste aanname uit de vorige alinea: we moeten een itemverzameling hebben met antwoorden van personen die aan een bepaald itemresponsmodel voldoen. Voordat we in de IRT gaan equivaleren moeten we eerst calibratieproblemen oplossen. Onder calibratie verstaan we het kiezen van een geschikt itemresponsmodel, het afnemen van data volgens een bepaald design, het schatten van de itemparameters en het toetsen op de geldigheid van het model. Calibratie is geen eenvoudige zaak en de problemen ermee in de praktijk moeten zeker niet onderschat worden. Een groot deel van de calibratie is reeds uitgebreid besproken in de hoofdstukken 4, 5 en 6. In paragraaf 8.3.1 zullen we een aantal aspecten nog eens de revue laten passeren. Indien de calibratie succesvol is afgesloten kunnen we de vaardigheid van de personen schatten op de vaardigheidsschaal. Dit onderwerp wordt in paragraaf 8.3.2 besproken. Hiermee zouden we IRT equivaleren kunnen afsluiten. Deze laatste twee paragrafen bespreken namelijk precies het equivaleren als we kunnen

werken met gecalibreerde itembanken: we zorgen voor een goede calibratie en de score op elke toets die we uit de bank samenstellen is automatisch geëquivaleerd middels vaardigheidsschattingen op de vaardigheidsschaal. De schaal waarop deze schattingen liggen kunnen we tenslotte nog transformeren naar een schaal die de gebruiker in staat stelt de resultaten goed te interpreteren. Aangezien dit laatste onderwerp uitgebreid wordt besproken in hoofdstuk 13, zullen we er hier verder geen aandacht aan besteden. De situatie waarin we met gecalibreerde itembanken kunnen werken zouden we actief equivaleren kunnen noemen: we stellen per definitie geëquivaleerde toetsen samen uit de itembank. In paragraaf 8.3.3 bespreken we een concreet voorbeeld van de opbouw en het werken met geëquivaleerde toetsen uit een itembank.

In de praktijk zijn er echter nog veel situaties waarin we passief moeten equivaleren: we beschikken over twee of meer toetsen waarvan de scores geëquivaleerd moeten worden. Van deze bestaande toetsen moet dan nagegaan worden of ze te calibreren zijn onder een IRT-model. Als er een passend IRT-model is gevonden, dan kan het soms nog een probleem zijn dat de resulterende schattingen op de vaardigheidsschaal komen te liggen en niet op een bestaande schaal voor de toets, bijvoorbeeld de ruwe scoreschaal. Een uitweg daarvoor kan bij IRT altijd worden gevonden via het zogenaamde ware score equivaleren, hetgeen we ook in paragraaf 8.3.2 zullen bespreken. Tenslotte zullen we in paragraaf 8.3.4 een mogelijke aanpak bespreken bij het equivaleren van bestaande toetsen als het gewenste IRT-model niet past.

8.3.1 Calibratie

Na de uitvoerige behandeling van de calibratie in de hoofdstukken 4, 5 en 6 zullen we ons hier beperken tot een aantal algemene overwegingen en factoren die direct gevolgen voor de praktijk van het equivaleren (kunnen) hebben. Welke factoren zijn dat nu precies? In de eerste plaats is (uiteraard) het gekozen itemresponsmodel van belang. Ten tweede kan het gebruikte design een rol spelen en ten derde moet er een methode gekozen worden waarmee de itemparameters geschat worden. Tenslotte besteden we ook nog enige aandacht aan het toetsen van het model. Al deze zaken impliceren keuzes en bovendien zijn deze keuzes niet onafhankelijk.

De keuze van het itemresponsmodel

Bij de keuze van het itemresponsmodel spelen vele factoren een rol. De toetsspecificatie, waarmee ondermeer bedoeld wordt het vaststellen van het doel van de toetsing en de keuze van het soort items, zie hoofdstuk 1, beperkt voor een groot deel de keuze uit de grote klasse van de bestaande IRT-modellen. Een paar voorbeelden: worden de items dichotoom dan wel polytoom gescoord; kan gokken een rol kan spelen, zoals bijvoorbeeld bij meerkeuze-items; is de te meten vaardigheid uni- of multidimensionaal. We zullen ons voorlopig beperken tot de unidimensionale modellen. Gegeven de toetsspecificatie moeten we binnen de beschikbare klasse een model kiezen. Een belangrijke overweging bij de keuze kan zijn, dat als we een model kiezen met voldoende statistieken voor de vaardigheidsparameter, dit automatisch leidt tot vaardigheidsparameterschatters die direct gekoppeld zijn aan de in de praktijk vaak gewenste (gewogen) ruwe scores op een toets. De keuze voor een bepaald itemresponsmodel heeft ook de belangrijke consequentie dat voor een deel de schattingsmethode reeds vastligt. Alleen als we kiezen voor een model met voldoende statistieken voor de vaardigheid hebben we, zoals uitvoerig betoogt in hoofdstuk 4 en 5, de voordelige eigenschappen van de CML-schattingsmethode ter beschikking en bovendien hebben we dan modeltoetsen met goede statistische eigenschappen. Een keuze voor bijvoorbeeld het drieparameter logistisch model, zie hoofdstuk 5, sluit de CML-schattingsmethode uit.

De eerste keuze voor een IRT-model wordt bepaald door het afwegen van theoretisch gewenste eigenschappen en praktische wensen en randvoorwaarden, echter deze keuze is soms slechts een voorlopige. Het is immers mogelijk dat tijdens de calibratie blijkt dat we met het gekozen model niet goed overweg kunnen en dat we een ander, vaak een ruimer, model moeten kiezen.

Het design

Het design is binnen de IRT een belangrijke factor. In hoofdstuk 6 hebben we gezien dat het design voor een gedeelte de schattingsmethode vastlegt. Bovendien is daar reeds uiteengezet dat om meer redenen de traditionele omweg van calibreren in volledige deeldesigns en het daarna op dezelfde schaal brengen van de itemparameters, soms het equivaleren van itemparameters genoemd, zo mogelijk vermeden dient te worden. Het schatten van de itemparameters dient in één calibratie plaats te vinden, ook als het design onvolledig is. Bovendien moeten we ons realiseren dat de keuze van een design vooral beperkt wordt door praktische randvoorwaarden, bijvoorbeeld in het geval dat

we twee bestaande toetsen gaan equivaleren. Alleen bij het actief equivaleren, het opbouwen van een itembank, staan doorgaans alle mogelijke designs ter beschikking.

Laten we de drie basisdesigns uit paragraaf 8.1.2 eens nader bekijken. Bij het eerste basisdesign, het single group design, zijn alle schattingsmethoden mogelijk. Bij het random group design, het tweede basisdesign, is er geen overlap tussen de items en ook niet tussen de personen. De extra aanname die bij dit design dan ook gemaakt dient te worden is dat de twee steekproeven uit één populatie getrokken zijn. Als we nu één vaardigheidsverdeling voor deze populatie aannemen, dan kunnen we met MML de itemparameters en ook de parameters van de vaardigheidsverdeling schatten. Merk op dat de CML schattingsprocedure bij het random group design nooit mogelijk is omdat dit design niet verbonden is. Het derde basisdesign, het ankertoetsdesign, heeft in zijn algemeenheid de ruimste toepassings- mogelijkheden en laat daarbij ook altijd nog een keuze voor de schattingsprocedure toe. Voor dit design is MML altijd mogelijk, en, als het model dit toelaat, CML ook.

Zoals eerder reeds opgemerkt is het ankertoetsdesign het enige basisdesign dat verticale equivalering mogelijk maakt. In dit verband moet er op gewezen worden dat in dat geval er wel speciale eisen aan de samenstelling van het anker moeten worden gesteld. We zullen dit met een voorbeeld toelichten. Als men toetsen calibreert die een onderwijstraject over een aantal jaren bestrijken en waarmee men de vorderingen van de leerlingen in kaart wil brengen, kan men niet met een vaste ankertoets werken. Vooruitgang op de ankertoets is namelijk bepalend voor de mogelijk te meten vooruitgang van de leerlingen over de jaren. In dit geval zal men per meetmoment ankers moeten kiezen die de vooruitgang kunnen weergeven. Zonder zorgvuldige analyse van het vaardigheidsdomein in de tijd en relevante keuzes voor de afnamemomenten kan het verticaal geëquivaleerde instrument mogelijk irrelevante veranderingen in de vaardigheid weergeven. In hoofdstuk 10 zal op dit onderwerp nog worden teruggekomen. Als algemene aanbeveling voor de samenstelling voor een ankertoets kan gesteld worden dat de inhoud ervan en ook de psyschometrische eigenschappen representatief moeten zijn voor de toetsen die het anker verbindt, zoals we ook al in paragraaf 8.2.2 zagen. Bij verticale equivalering impliceert dit dus ook een goede spreiding van de items qua moeilijkheid.

Toetsing van het model

Daar de modeltoetsing reeds uitgebreid behandeld is in hoofdstuk 4, volstaan we hier met het maken van een tweetal opmerkingen. De eerste opmerking betreft de calibratie

voor het verticaal equivaleren. Om verticaal te kunnen equivaleren zal, daar de vaardigheids- verdelingen flink kunnen verschillen, de verbondenheid uit de items moeten komen. Dat wil dus zeggen dat de ankeritems door personen met flink uiteenlopende vaardigheden gemaakt zullen gaan worden. Een belangrijke vraag in dit verband is dan: meten deze items wel hetzelfde in de verschillende populaties? Naast de gebruikelijke toetsing van het IRT-model, zullen we hierop speciaal moeten toetsen. Hoe hierop getoetst moet worden is het onderwerp van hoofdstuk 9, dat het onderwerp itemonzuiverheid behandeld. We zullen hier dan ook niet verder op ingaan.

De tweede opmerking heeft te maken met slecht passende items. Bij de calibratie zullen er, zoals de ervaring leert, naar alle waarschijnlijkheid items verwijderd moeten worden die om de een of andere reden niet aan het gekozen itemresponsmodel voldoen. Als de calibratie dient om een itembank te construeren, dat wil zeggen om een verzameling van items te vinden die op dezelfde schaal liggen, dan is er geen probleem. Tenminste, als de domeinomschrijving van de overgebleven items nog voldoende dekking geeft zodat we nog steeds hetzelfde meten. Anders is het als de equivalering plaats dient te vinden op bestaande toetsen, eerder passieve equivalering genoemd. We kunnen de equivalering dan uitvoeren met de overgebleven items. Een nadeel hiervan kan zijn dat de leerlingen slechts op een gedeelte van de werkelijk gemaakte toets worden beoordeeld. Dit kan problematisch en oneerlijk zijn, denk hierbij bijvoorbeeld aan de eindexamens. In dat geval zullen we óf een itemrespons-model moeten kiezen waarbij géén items meer verwijderd hoeven te worden óf we zullen moeten equivaleren met behulp van de KTT.

8.3.2 Verschillende vormen van equivalering in de itemresponstheorie

Binnen de IRT zijn er, net zoals in de KTT, in principe, twee methoden in gebruik om te equivaleren. De eerste methode, die het vaakst wordt gebruikt, is het equivaleren via het schatten van de vaardigheid. Hierbij wordt voor elke persoon op basis van zijn antwoord-patroon een schatting $\hat{\theta}$ van zijn of haar latente vaardigheid θ berekend. Deze schattingen zijn dan gelijk geëquivaleerd, daar ze op dezelfde schaal liggen. De tweede methode, die met name in de Amerikaanse literatuur veel wordt besproken, zie bijvoorbeeld Lord (1980), is het ware score equivaleren. Deze methode, die met name gebruikt wordt bij het equivaleren van bestaande toetsen, gebruikt ook schattingen van θ en transformeert deze naar een schaal die past bij de oorspronkelijke ruwe (en ware) score schaal van de toets. Alvorens deze methoden te bespreken merken we op dat beide methoden ervan uitgaan dat calibratie van alle items succesvol is verlopen. We

beschikken dan dus over schattingen van de itemparameters, die daarna als vast verondersteld worden. Bij het berekenen van de vaardigheidsschattingen gaan we er dan eigenlijk ten onrechte van uit dat de itemparameters geen schattingsfout hebben. Over het precieze effect van deze benadering is nog slechts weinig bekend. Dit effect wordt uiteraard geringer naarmate de schattingsfouten van de itemparameters kleiner zijn. De grootte van de steekproef en het afnamedesign zijn hiervoor bepalend.

Het schatten van de vaardigheid

In hoofdstuk 4 zijn drie methoden voor het schatten van de vaardigheid behandeld te weten de ML, WML en de bayesiaanse schattingsmethode EAP. De eigenschappen en respectieve-lijke voor- en nadelen van deze methoden zijn daar reeds uitgebreid besproken. Een voorbeeld met een vergelijking van schattingen met deze methoden staat in tabel 4.13. Hier volstaan we met een aantal opmerkingen over de keuze van een schatter voor de vaardigheid in relatie tot de schattingsmethode die bij de calibratie is gevolgd. Voor de keuze van een methode voor het schatten van de vaardigheid is het van belang of het itemresponsmodel wel of geen voldoende statistieken voor de vaardigheid heeft. In modellen zonder voldoende statistieken voor de vaardigheidsparameter moet de calibratie, als we de JML-methode vanwege het niet consistent zijn van de itemparameterschatters buiten beschouwing laten, altijd met de MML of andere in dit boek niet besproken bayesiaanse methoden worden uitgevoerd. Het is een gemeenschappelijk kenmerk van deze methoden dat het gebruikte itemresponsmodel wordt aangevuld met een (of meer) verdeling(en) voor de vaardigheid. Laten we even aannemen dat we beschikken over slechts één populatie. De aanname van een vaardigheidsverdeling voor deze populatie betekent eigenlijk dat de vaardigheid van de personen niet meer vast of fixed is, maar random, dat wil zeggen getrokken uit een bepaalde, al dan niet compleet gespecificeerde, vaardigheidsverdeling. Tijdens de calibratie moeten dan zowel de itemparameters als de (eventuele) parameters van de vaardigheidsverdeling gezamenlijk geschat worden. Het model geldt dus alleen onder de extra aanname van deze vaardigheidsverdeling. Aan de ene kant kunnen we nu stellen dat we bij de schatting van de vaardigheid van individuele personen rekening dienen te houden met het feit dat ze getrokken zijn uit een bepaalde populatie met een onderliggende verdeling. Maar dit betekent dat we de vaardigheid met een bayesiaanse methode moeten bepalen. De EAP-methode komt dan in aanmerking. Als we namelijk bij de schatting van de vaardigheidsparameter géén gebruik maken van deze onderliggende verdeling, dan gebruiken we niet alle beschikbare informatie, zodat deze

schatting statistisch niet optimaal kan zijn. Aan de andere kant kunnen we ook stellen dat de calibratie alleen maar dient om de itemparameters te schatten. De aanname van een vaardigheidsverdeling was alleen maar noodzakelijk om de schaal vast te leggen. Bij de schatting van de vaardigheid hoeven we hier dus geen rekening meer mee te houden. In de praktijk wordt bijna altijd gekozen voor de tweede optie. Er wordt dan dus géén rekening gehouden met de onderliggende vaardigheidsverdeling en het informatieverlies wordt op de koop toe genomen. In concreto betekent dit dat de vaardigheidsparameter θ gewoon met de ML- of WML-methode geschat wordt. In modellen met voldoende statistieken voor de vaardigheid kan de calibratie uitgevoerd worden met zowel CML als MML. Als we gecalibreerd hebben met CML, een methode die steekproefonafhankelijk is, kunnen we de vaardigheid schatten met de ML- of WML-methode. Als de calibratie met MML is geschied, geldt hetzelfde als in modellen zonder voldoende statistieken, zoals hiervoor uiteengezet. Ook dan worden ML- of WML-schattingen voor de vaardigheid gebruikt.

Als we bij de schatting van de vaardigheidsparameter géén gebruik (wensen te) maken van populatiegegevens, dan gaat, voor elk itemresponsmodel, de voorkeur uit naar WML-schatters, daar deze, bij benadering, zuivere schatters van de vaardigheid opleveren (zie hoofdstuk 4). Zoals bekend zal de nauwkeurigheid van deze schatters (standaardfout kleiner) en dus van de equivalering toenemen naarmate de moeilijkheid van de toets dichter bij de te schatten vaardigheid ligt.

Ware score equivalering

Bij het equivaleren van bestaande toetsen, en soms ook als men toetsen samenstelt uit een itembank, wenst men na equivalering te rapporteren naar de gebruiker op de (eventueel nog te transformeren) ruwe score schaal, dat wil zeggen het aantal items goed. Schattingen op de vaardigheidsschaal hebben daar niet altijd een direct verband mee. Als we toetsen beschouwen met dichotome items en als IRT-model het twee- of drieparametermodel, dan levert elk verschillend antwoordpatroon een verschillende schatting van de vaardigheid op. Ter illustratie beschouwen we een voorbeeld. We hebben de gegevens geanalyseerd van een subtoets van de zogenaamde Scholastic Apptitude Test (LSAT-6), die vermeld staan in Mislevy en Bock (1986). Deze subtoets bestaat uit vijf items. Met de antwoorden van 1000 personen werd een calibratie uitgevoerd met het tweeparametermodel en met het Raschmodel. Vervolgens werden de vaardigheden van deze personen geschat met de EAP-methode. Een deel van de

resultaten staat in tabel 8.1, en wel de EAP-schattingen voor personen die 3 of meer scoorden op deze toets.

Tabel 8.1
EAP-vaardigheidschattingen tweeparametermodel en Raschmodel LSAT-6

Tweeparametermodel				Raschmodel		
Patroon	score	aantal	EAP	score	aantal	EAP
00111	3	4	-.314	3	237	-.331
01011	3	16	-.395			
01101	3	3	-.296			
01110	3	2	-.275			
10011	3	81	-.366			
10101	3	28	-.266			
10110	3	15	-.245			
11001	3	56	-.347			
11010	3	21	-.326			
11100	3	11	-.226			
01111	4	15	.062	4	357	.063
10111	4	80	.093			
11011	4	173	.008			
11101	4	61	.112			
11110	4	28	.134			
11111	5	298	.498	5	298	.477

We zien dat, als we het tweeparametermodel gebruiken, voor elk antwoordpatroon een andere schatting voor de vaardigheid volgt. Dit in tegenstelling tot als we het Raschmodel gebruiken: in dat model is immers de somscore een voldoende statistiek voor θ, en krijgen we alleen voor verschillende somscores verschillende vaardigheids-schattingen. Voor de volledigheid zij vermeld dat de schattingen in tabel 8.1 gerapporteerd staan op een schaal, die genormeerd is op de vaardigheidsverdeling. Deze verdeling heeft een gemiddelde van 0 en een standaarddeviatie van .075.

Bij het tweeparametermodel, en in het algemeen met modellen die geen voldoende statistiek voor θ hebben, is er dus geen directe relatie tussen de geschatte vaardigheden en de (eventueel gewogen) ruwe score schaal. Deze schattingen hebben dus ook geen

directe relatie met de ruwe scores van de te equivaleren toetsen. Als men de te equivaleren toetsen op de ruwe score schaal zou willen rapporteren, komt men met de geschatte vaardigheden niet verder. Een werkwijze die men dan kan toepassen is ware score equivalering, die als volgt werkt.

Men definieert de ware score op een toets, vergelijkbaar met de ware score in de KTT, als de verwachtingswaarde van de ruwe score:

$$\tau_X = \mathscr{E}(X) = \mathscr{E}(\sum_{i \in X} X_i) = \sum_{i \in X} \mathscr{E}(X_i) = \sum_{i \in X} P_i(\theta), \tag{8.22}$$

waarbij $P_i(\theta)$ de kans op een goed antwoord onder het gebruikte IRT-model is. Het is eenvoudig in te zien, dat bij dichotome items de ware score precies het bereik heeft van de ruwe score schaal. De ware score (8.22) als functie van θ beschouwd, wordt ook wel de toetskarakteristieke functie genoemd en is de som van de itemresponsfuncties van de items waaruit de toets bestaat. Een schatting van de ware score van een persoon op een toets verkrijgt men door het evalueren van (8.22) in het punt van de schatting van de persoon op de vaardigheidsschaal $\hat{\theta}: \hat{\tau}_X = \Sigma_{i \in X} P_i(\hat{\theta})$.

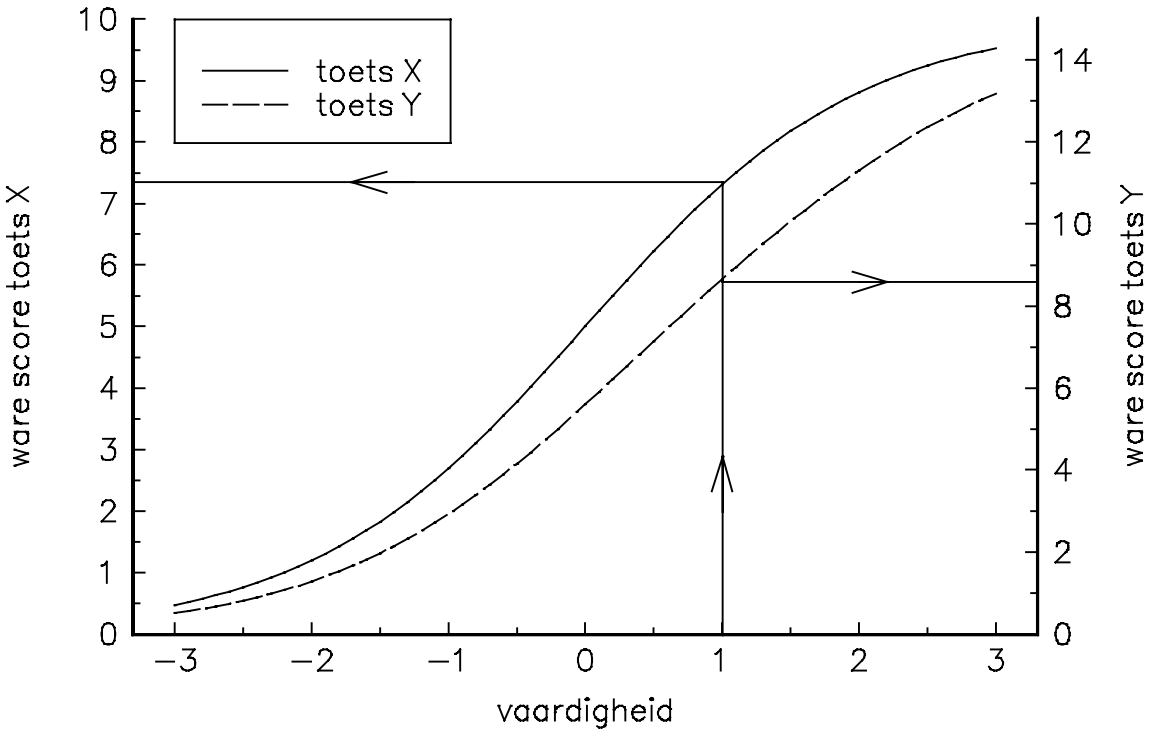
Als we nu twee toetsen X en Y hebben die gecalibreerd zijn onder een IRT-model, dan kan men de geschatte ware scores op beide toetsen die horen bij een bepaalde θ als geëquivaleerde scores beschouwen. Voor de te equivaleren toetsen X en Y zijn de ware scores als functie van θ gegeven door

$$\begin{aligned} \tau_X &= \sum_{i \in X} P_i(\theta) \quad \text{en} \\ \tau_Y &= \sum_{j \in Y} P_j(\theta). \end{aligned} \tag{8.23}$$

Voor elke θ en dus ook voor elke schatting van $\hat{\theta}$ van θ zijn dan de ware scores en dus ook de geschatte ware scores $\hat{\tau}_X$ en $\hat{\tau}_Y$ equivalent. Met een voorbeeld zullen we dit toelichten. In figuur 8.7 staan de toetskarakteristieke functies van toets X, bestaande uit 10 items, en toets Y, die uit 15 items bestaat. Als voorbeeld is aangegeven dat bij $\theta = 1$ de ware score op toets X gelijk is aan 7.35 is en voor toets Y gelijk aan 8.29, de equivalente scores op deze toetsen bij deze waarde van θ. Voor elke θ kunnen we op deze manier equivalente scores op de toetsen vinden.

In de praktijk gebruikt men ware score equivalering ook nog wel eens op de volgende manier. Stel dat men toets Y wil equivaleren met een vroegere versie toets X en men wil weten wat de equivalente score is van een ruwe geobserveerde score op toets Y op de ruwe score schaal van toets X. Men wil dan dus ruwe geobserveerde scores equivaleren. In plaats van de ware score op toets Y gebruikt men dan de

geobserveerde ruwe score en zoekt daarbij de bijpassende score op de schaal van toets X. Als voorbeeld in figuur 8.7 vinden we dan bij een score 6 op toets Y een score van 5.2 op toets X. Alhoewel er theoretisch geen recht- vaardiging is voor het op deze manier equivaleren van geobserveerde scores, blijkt het in de praktijk redelijke resultaten op te leveren (Lord & Wingersky, 1983). Merk op dat voor het Raschmodel ware score IRT equivalering identiek is aan deze vorm van geobserveerde score IRT equivalering. Bij elke geobserveerde ruwe score hoort in het Raschmodel immers maar één schatting $\hat{\theta}$.



Figuur 8.7
Ware score equivalering van twee toetsen X en Y

8.3.3 Equivaleren met behulp van een itembank

In deze paragraaf behandelen we een voorbeeld van de opbouw van een itembank, dat wil zeggen het calibreren en het samenstellen van geëquivaleerde toetsen uit de bank. Dit concrete voorbeeld betreft de schaal vorderingen in spellingvaardigheid (SVS; Van den Bosch, Gillijns, Krom & Moelands, 1991). De SVS is een instrument om (vorderingen in) spellingvaardigheid te meten voor de groepen drie en vier van het

basisonderwijs. Na proefafnames zijn er negen verschillende modules samengesteld, elk van ongeveer 20 items. Daarna zijn deze modules afgenomen bij een landelijke steekproef middels het (longitudinale) design zoals gegeven in figuur 8.8. Boekje 1 bijvoorbeeld, dat is samengesteld uit de modules 1 en 2, is afgenomen op tijdstip m3 (medio groep 3) bij sag a. Een sag is een school afname groep en dient ter vereenvoudiging van de afname procedure; elke school in een sag maakt per afnametijdstip één boekje. Merk op dat binnen elk tijdstip het design verbonden is. Bovendien is het design over de tijdstippen heen verbonden en is het afnameschema zo geconstrueerd dat geen enkele leerling twee maal dezelfde module maakt, waardoor herinneringseffecten vermeden worden. Module 3 bijvoorbeeld, is op het eerste tijdstip (m3) gemaakt door leerlingen uit sag b en sag c, en een tijdstip later (e3, eind groep 3) door leerlingen uit sag a. Of, andersom bekeken, leerlingen

			Module								
			1	2	3	4	5	6	7	8	9
boekje	sag	tijd									
1	a	m3	■	■							
2	b			■	■						
3	c		■		■						
4	a	e3			■	■					
5	b					■	■				
6	c			■		■					
7	a	m4					■	■			
8	b							■	■		
9	c						■		■		
10	a	e4							■	■	
11	b									■	■
12	c							■		■	

Figuur 8.8
Calibratiedesign Spellingvaardigheid

uit sag a maken op de verschillende afnametijdstippen achtereenvolgens de modules 1+2, 3+4, 5+6 en 7+8, nooit dezelfde dus. Merk bovendien op dat een module die het design voor twee aanliggende tijdstippen verbindt, alleen op die twee tijdstippen is ingezet. Er is dus geen vast anker gebruikt (zie ook paragraaf 8.3.1). Omdat het voor rapportage- en onderwijskundige doeleinden het noodzakelijk was om over genoeg

gegevens omtrent de spelling van allochtone leerlingen te beschikken, zijn binnen elke sag de scholen met relatief veel allochtone leerlingen oververtegenwoordigd. Dit heeft als belangrijke consequentie dat voor een willekeurig gekozen tijdstip de steekproef niet meer representatief is voor de populatie op dat tijdstip. Bepaalde groepen zijn oververtegenwoordigd en de leerlingen zijn ook nog eens in clusters (scholen) getrokken. Uit de proefafname was bovendien bekend dat een goede beschrijving van de antwoorden op de items mogelijk was als we gebruik maakten van het OPLM. Om dezelfde reden als in paragraaf 7.1, geven we dan de voorkeur aan een calibratie met de CML-methode, deze methode is immers steekproefonafhankelijk. Alle (173) afgenomen items bleken op de SVS schaal te passen. In deze schaal zitten dus bijvoorbeeld geen items meer die tijdstip-onzuiverheid vertonen. Voor elke leerling die een bepaald boekje gemaakt heeft, kunnen we nu aan de hand van zijn toetsscore een schatting van zijn vaardig- heid maken. Deze vaardigheidsschattingen gebruiken we op verschillende manieren. De eerste, en meest belangrijke, is voor de bepaling van referentiegegevens. Deze referentie- gegevens worden per tijdstip zowel voor de totale populatie als ook voor de subpopulatie van allochtonen bepaald; de procedure hiervoor staat beschreven in hoofdstuk 10. Merk op dat bij de bepaling van de referentiegegevens op populatieniveau, er rekening mee gehouden dient te worden dat de allochtonen in de steekproef oververtegenwoordigd waren. Bovendien worden de vaardigheidsschattingen van de leerlingen naar de scholen die aan de calibratie hebben deelgenomen gerapporteerd.

Nadat de itembank SVS was geconstrueerd, zijn er voor elk afnametijdstip modules op maat samengesteld. Hiermee kan de leerkracht een leerling een toets voorleggen die meer toegespitst is op zijn of haar vaardigheid. De minder goede leerling krijgt dan een makkelijke en de goede leerling een moeilijke module. De belangrijkste reden voor dit toetsen op maat is dat de schattingsfouten van de vaardigheid flink kleiner worden. Bij WML, bijvoorbeeld, worden de schattingsfouten gemiddeld ongeveer dertig procent kleiner. Omdat de itembank gecalibreerd is, zijn de vaardigheidsschattingen op de verschillende modules gelijk geëquivaleerd. Bovendien kunnen deze geëquivaleerde scores direct gerelateerd worden aan de referentiegegevens: we kunnen nu immers de relatieve positie van de leerling in de betrokken populatie bepalen (zie ook hoofdstuk 10). Ook kan de vaardigheid van de leerling gerelateerd worden aan relevante onderwijskundige criteria (Van den Bosch e.a., 1991).

Een laatste opmerking. Omdat we werken met OPLM, zullen voor een juiste afspiegeling van de vaardigheid gewogen scores gebruikt moeten worden. In de praktijk wordt er door de leerkracht, voor wie de SVS als hulpmiddel dient, voornamelijk gebruik gemaakt van ongewogen (ruwe) scores. Er is daarom dan ook een procedure

ontwikkeld die aan dit probleem tegemoet komt. We zullen hier verder echter niet op ingaan.

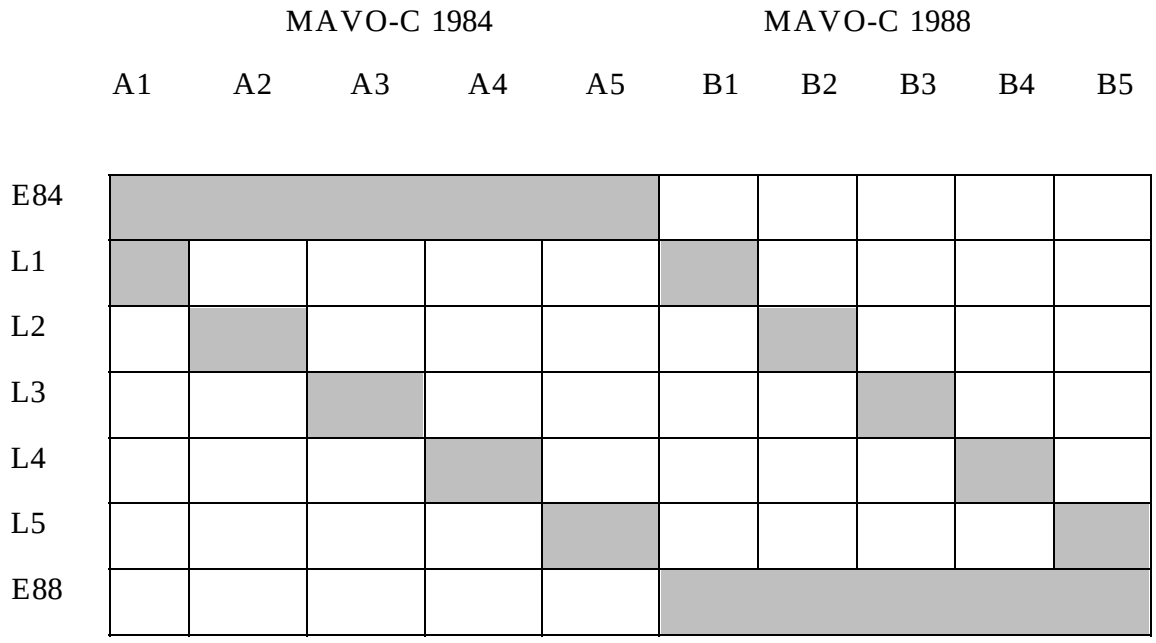
8.3.4 Quasi-multidimensionaal IRT-equivaleren

Zoals reeds in de inleiding is opgemerkt worden elk jaar de twee tijdvakken van een aantal centraal schriftelijke examens geëquivaleerd. Maar hoe zit dat nu met de examens over de jaren heen? Is het eindexamen van 1992, zeg, vergelijkbaar met dat van 1993? Dit is niet alleen een moeilijk maar ook, zeker voor belanghebbenden zoals leerlingen en onderwijsgevenden, een belangrijk probleem. In het vervolg zullen we ons voor het gemak beperken tot examens waarbij de items dichotoom gescoord worden. Een eerste opmerking die hier van belang is, betreft de scoringsregel die bij de examens gehanteerd wordt. Bij de examens moet de behaalde score een functie zijn van het aantal goed gemaakte opgaven. Bovendien moet elke opgave 'even zwaar' meetellen in het eindresultaat. Dit heeft als belangrijkste consequentie dat er een beperking op het te kiezen IRT-model ligt: alleen modellen met gelijke discriminatie-parameters komen in aanmerking. Het enige model dat dan nog over blijft is het Raschmodel. Voor de calibratie-methode komen dan zowel MML als CML in aanmerking. Bovendien zijn we bij examens behalve in equivalente scores over verschillende jaren ook in het slagingspercentage geïnteresseerd. Dit betekent dat we graag willen weten hoeveel procent van de kandidaten uit 1993 zou geslaagd zijn als ze het examen van 1992 gemaakt hadden. Daar dit laatste een kenmerk van de populatie is, ligt het voor de hand om de calibratie uit te voeren met MML.

Hoe de equivalering van twee examens uitgevoerd kan worden, zullen we demonstreren aan de hand van een voorbeeld. Als voorbeeld nemen we de examens frans van de jaren 1984 en 1988 voor MAVO-C. Eerst zijn beide examens in vijf delen geknipt. Voor het 1984 examen noemen we deze delen A1 tot A5 en voor het examen van 1988 duiden we deze delen aan met B1 tot B5. Vervolgens zijn deze delen, net na de afname van het examen in 1988, volgens het design in figuur 8.8 afgenomen bij een steekproef van leerlingen uit klas 3 van het VWO. De groepen L1 tot L5, allen uit klas 3 van het VWO, maken dus steeds een gedeelte van het 1984 en een gedeelte van het 1988 examen. Het ligt namelijk in de lijn der verwachting dat de vaardigheid van deze leerlingen vergelijkbaar is met de vaardigheid van de eindexamen kandidaten in MAVO-C (Glas, 1989).

Nu valt het niet te verwachten valt dat beide examens op een unidimensionale schaal liggen, omdat examens immers van de kandidaten diverse 'vaardigheden' vragen. Dit betekent dan ook dat het Raschmodel voor de totale itemverzameling naar verwachting niet zal passen, wat in werkelijkheid ook zo bleek te zijn. Daarom is gezocht naar een multi-dimensionale oplossing voor het equivaleerprobleem. Het bleek namelijk dat de totale itemverzameling op te splitsen was in een aantal subschalen die alle aan het Raschmodel voldeden. De gebruikte procedure om tot deze subschalen te komen werkt als volgt. Eerst moeten de vaardigheids-verdelingen gespecificeerd worden. Voor elk van de drie onderscheiden groepen, te weten de examen kandidaten van 1984 (E84), leerlingen uit klas 3 van het VWO (L1-L5) en de examen kandidaten van 1988 (E88) nemen we een normale verdeling aan. De schaal wordt vastgelegd door het gemiddelde van de vaardigheidsverdeling van de 1984 examinandi gelijk aan nul te stellen.

We gaan nu de eerste subschaal zoeken. Dit doen we door uit de totale set van items die items te verwijderen die op basis van de itemgerichte R_{1m} toets niet blijken te passen. Dit doen we net zo lang totdat er een schaal gevonden is. Bij deze schaal kunnen dus geen items meer verwijderd worden op basis van de R_{1m} toets. Deze unidimensionale Raschschaal noemen we subschaal 1. Vervolgens zoeken we de tweede subschaal op precies dezelfde



Figuur 8.8
Equivaleerdesign MAVO-eindexamen

manier als hierboven uit de overgebleven items, dat wil zegen uit de totale set van items behalve de items uit subschaal 1. Uiteindelijk werden er drie subschalen gevonden en bleken slechts vier items (alle uit 1984) van de in totaal 100 items op geen enkele subschaal te passen. Het blijkt dus dat we zelfs met dit multidimensionale itemresponsmodel niet alle items kunnen schalen. We zouden dus nu eigenlijk een ruimer IRT model moeten kiezen. Dit is mogelijk, daar er voor dit soort items modellen bestaan waarbij een item op meerdere vaardigheidsdimensies laadt, zie bijvoorbeeld paragraaf 5.5. Voor de beschrijving van dit voorbeeld zullen we echter aannemen dat de calibratie met succes is afgesloten, de vier niet passende items ten spijt. We beschikken nu over drie subschalen met per subschaal drie vaardigheidsverdelingen, voor elk van de onderscheiden groepen leerlingen één. De linking groepen, dat wil zeggen de leerlingen uit klas 3 van het VWO, zijn nu verder niet meer van belang, daar deze alleen maar dienden om het design te verbinden.

Uiteindelijk hebben we op deze manier nu precies een model zoals beschreven in paragraaf 5.5. Merk op dat elk examen uit drie subschalen bestaat, een leerling heeft op elke subschaal een vaardigheid. Laten we eens aannemen dat een leerlinge 43 items goed beantwoord heeft van het 1984 examen. Deze score van 43 kan op zeer veel verschillende manieren tot stand gekomen zijn. De leerlinge kan bijvoorbeeld van de eerste subschaal 20 items goed hebben, van de tweede 17 en van de laatste subschaal 6. Bij deze combinatie horen uiteraard drie vaardigheidsschattingen, op elke subschaal een. Omdat we bij de examens niet op de vaardigheidsschaal werken, moeten we dus deze vaardigheidschattingen gebruiken om op elke subschaal een equivalente score op dezelfde subschaal van 1988 examen te zoeken. Of, met andere woorden, op elke subschaal passen we ware score equivalering toe. Tenslotte berekenen we de equivalente score van deze leerlinge op het totale 1988 examen door de som van de drie geëquivaleerde scores (op de subschalen) te nemen. Het is eenvoudig in te zien dat voor een andere leerling met 43 items goed in 1984, best een andere geëquivaleerde score in 1988 gevonden kan worden.

Een van de belangrijkste waarden bij een examen is de cesuur, dat wil zeggen de score, waar de grens tussen een onvoldoende en een voldoende ligt. We kunnen nu de cesuur voor het 1988 examen berekenen op grond van de populatie uit 1984. Hiermee kunnen we dan gelijk de vraag beantwoorden hoeveel kandidaten uit 1984 voor het 1988 examen geslaagd zouden zijn. Daarvoor schatten we eerst voor elke 1984 leerling de vaardigheidsparameters $\hat{\theta}_{84q}$, $q = 1,...,3$, waarbij q de subschaal weergeeft. De somscore op het examen van 1988, r^*_{88}, wordt vervolgens geschat door

$$r^*_{88} = \sum_{q=1}^{3} \sum_{i \in I_q} \mathscr{E}(X_i \mid \hat{\theta}_{84q}, \hat{\delta}_q), \tag{8.23}$$

waarbij δ_q de itemparameters van het 1988 examen zijn en I_q die items die op subschaal q van het 1988 examen liggen. Bovenstaande formule geeft dus de verwachting van de score van een 1984 examinandus op het 1988 examen. Als we voor elke leerling (8.23) berekenen, en de cesuur van 1988 toepassen, kunnen we dus gelijk vaststellen hoeveel procent van de 1984 populatie in 1988 geslaagd zou zijn.

8.4 De kwaliteit van de equivaleermethoden vergeleken

Bij de beschrijving van de equivaleermethoden in dit hoofdstuk zijn soms voor- en nadelen genoemd. Dit is één bron om de kwaliteit van de methoden te vergelijken. De tweede is om terug te grijpen op de zeer omvangrijke psychometrische literatuur die de laatste jaren is verschenen en nog verschijnt over studies die tot doel hadden equivaleermethoden te vergelijken. Het is in dit verband niet zinvol om uitvoerig op deze studies in te gaan. Op de eerste plaats heeft dit te maken met de enorme hoeveelheid artikelen die over het onderwerp verschijnen; het volledig bespreken zou zeer veel tijd kosten. In de tweede plaats zijn deze studies vaak zeer specifiek toegespitst op één bepaald aspect van één equivaleermethode, zodat ze slechts geringe generalisatiemogelijkheden hebben. In de derde plaats is de kwaliteit van de artikelen vaak matig. De voorwaarden en aannamen waaronder een bepaalde techniek geldig is, worden zelden expliciet genoemd. Een veel voorkomende fout is bijvoorbeeld dat de kwaliteit van IRT equivalering als slecht wordt beoordeeld, terwijl het gehanteerde model niet past. In dit geval kan echter geen oordeel over de kwaliteit plaatsvinden, daar de equivalering slechts bij modelpassing kan worden uitgevoerd.

Een integratie van beide bronnen leidt tot de volgende conclusies. De eerste en belang- rijkste conclusie is dat equivaleren met behulp van de IRT in het algemeen de voorkeur heeft boven equivaleren met behulp van de KTT. Uiteraard moet dan bij het gebruik van een bepaald itemresponsmodel allereerst de modelgeldigheid nagegaan worden. De strenge eisen die bij de modeltoetsing worden opgelegd hebben als rechtstreeks gevolg dat de equivalering eenvoudig wordt. Als we over IRT equivaleren praten, zullen we steeds aannemen dat de calibratie met succes is afgesloten. Indien het gekozen itemresponsmodel echter niet past, en een ruimer model ook geen oplossing geeft, dan kunnen we altijd terugvallen op de KTT, welke immers minder stringente eisen aan de data stelt. In dat geval moeten we er ons echter wel bewust van zijn dat we nu meestal enkele niet toetsbare aannames en vooronderstellingen moeten maken.

De tweede conclusie is dat IRT equivaleermethoden eerder werken naarmate het aantal parameters groter is, omdat dan de modellen eerder passen. Het blijkt echter,

dat er voor itemresponsmodellen met veel parameters, zoals bijvoorbeeld het 3PL, geen goede toetsen beschikbaar zijn, behalve hele strenge toetsen. Denk hierbij bijvoorbeeld aan toetsen die met behulp van kruisvalidatie-technieken geconstrueerd kunnen worden (zie ook hoofdstuk 5).

De derde conclusie slaat alleen op equivaleermethoden binnen de KTT. Hier blijkt dat bij het gebruik van het single group design of het random group design alle equivaleermethoden, binnen praktisch relevante marges, overeen komen. Bij het ankertoetsdesign gelden ongeveer dezelfde conclusies, mits het anker aan de in dit hoofdstuk reeds besproken (psychometrische) voorwaarden voldoet en het aantal ankeritems groot genoeg is.

Tenslotte nog een laatste opmerking. In dit gehele hoofdstuk zijn schattingsfouten doorgaans buiten beschouwing gelaten. Enerzijds is dit gebeurt om het niet nodeloos ingewikkeld te maken, anderzijds omdat er slechts weinig analytische resultaten bekend zijn. In de literatuur worden de equivaleerfouten meestal gekarakteriseerd als systematisch en random. De systematische fouten zijn dan het rechtstreekse gevolg van het schenden van de assumpties. Als we bijvoorbeeld het random group design bekijken, dan kan het zo zijn dat de verschillende groepen niet vergelijkbaar zijn. Het moge duidelijk zijn dat systematische fouten ten alle tijden zoveel mogelijk vermeden dienen te worden. Daaruit volgt logischerwijs dat de assumpties op de een of andere manier getoetst moeten worden. Hoe deze assumpties, indien mogelijk, getoetst kunnen worden is beschreven bij de bespreking van de verschillende methoden. Merk op dat het toetsen van de assumpties voornamelijk een groot probleem is bij equivaleren in de KTT. Omdat we in de praktijk altijd met steekproeven werken waarmee populatie kenmerken geschat moeten worden, zullen we altijd statistische fouten maken (random equivaleerfouten). Om deze zo klein mogelijk te maken is het een eerste vereiste dat de steekproef voldoende groot is. Bovendien verdient het uiteraard aanbeveling om de steekproef af te stemmen op de te equivaleren toetsen. Dit laatste is voornamelijk een groot voordeel bij equivaleren in de IRT, bijvoorbeeld bij 'toetsen op maat'. Voor meer informatie omtrent (statistische) schattingsfouten als we equivaleren in de KTT, verwijzen we naar Braun en Holland (1982), Lord (1950) en Angoff (1971).