Equivaleren op basis van aanvullende data

1. Inleiding
2. Een model voor het equivaleren van examens
3. Het schatten van parameters en scores
4. Gegevensverzamelingen voor het equivaleren van examens

1. Inleiding

In het document 'Equivaleren en equivaleringsprocedures' is uitgelegd dat twee equivalente examens tot vergelijkbare uitspraken over één bepaalde populatie zouden moeten komen. De basis voor de equivalering vormde het idee van één populatie die beide examens had afgelegd. In de werkelijkheid zal de examenpopulatie nooit bereid zijn om alleen ten behoeve van de normering nog een extra examen af te leggen. Door de aanvullende dataverzameling waarbij referentie-items en examenitems gecombineerd worden afgenomen, kunnen we uitspraken doen over verschillen in moeilijkheidsgraad van de opgaven uit beide examens. Als we die verschillen kennen, kunnen we ook schatten wat een andere populatie op de makkelijkere of moeilijkere opgaven zou scoren. Zo kunnen we ook een beeld krijgen van eventuele verschillen in vaardigheidsverdeling van beide populaties. De volgende toelichting zet uiteen welke principes en methoden gebruikt worden om het College voor Examens (CvE) de voor de normering benodigde gegevens te verschaffen. We werken daarbij aan de hand van een statistisch model.

2. Een model voor het equivaleren van examens

Om toetsen te kunnen equivaleren moet men beschikken over een model voor het beschrijven van de toetsresultaten. We zullen daarom beginnen met het introduceren van een zeer eenvoudig, en daarom ook zeer onrealistisch model. In tabel 1 zijn de resultaten van 6 leerlingen op 6 opgaven weergegeven. In de eerste rij van de tabel staan de scores van de eerste leerling, in de eerste kolom staan de scores op het eerste item, etc. Ontbrekende scores zijn aangegeven met een ster. Dus leerling 4 heeft alleen de opgaven 1, 3 en 6 gemaakt.

***Tabel 1**: scores van 6 leerlingen op 6 open vragen*

leer-	item					
ling	1	2	3	4	5	6
1	8	7	4	9	5	4
2	7	6	3	8	4	3
3	6	5	2	7	3	2
4	5	*	1	*	*	1
5	*	6	*	*	4	*
6	*	*	*	9	*	*

Tabel 1 is zo geconstrueerd dat de scores van de leerlingen door een zeer eenvoudig model kunnen worden beschreven. In tabel 2 is ieder antwoord geschreven als de som van twee factoren; één factor hangt samen met de vaardigheid van de leerling en de andere factor hangt samen met de moeilijkheidsgraad van de opgave. Deze twee factoren (of effecten) noemt men respectievelijk een persoonsparameter (θ_i, thêta) en een itemparameter (δ_j, delta). Elke score kunnen we zo herschrijven tot $X_{ij}=\theta_i+\delta_j$. Op deze wijze zijn de scores uit tabel 1 herschreven tot samengestelde scores in tabel 2.

***Tabel 2**: item- en leerlingbijdrage aan scores van 6 leerlingen op 6 open vragen*

leer-	item						
ling	1	2	3	4	5	6	θ
1	3+5	3+4	3+1	3+6	3+2	3+1	3
2	2+5	2+4	2+1	2+6	2+2	2+1	2
3	1+5	1+4	1+1	1+6	1+2	1+1	1
4	0+5	*	0+1	*	*	0+1	9
5	*	2+4	*	*	2+2	*	2
6	*	*	*	2+6	*	*	3
δ	5	4	1	6	2	1	

Door gebruik te maken van dit model kunnen we een voorspelling doen over de ontbrekende antwoorden. Zo voorspellen we volgens dit model dat leerling 5 op item 1 een score 7 (=2+5) haalt, omdat de persoonsparameter van leerling 5 gelijk is aan 2, en de itemparameter van item 1 gelijk is aan 5.

Het hierboven beschreven model is echter onbevredigend omdat het in de praktijk nooit bij toetsresultaten zal passen. In de eerste plaats is een eenvoudige optelling in de werkelijkheid nooit toereikend om de gegeven scores te beschrijven. Verder suggereert het model, dat we gegeven de persoons- en itemparameters exact weten welke score een leerling haalt. In iedere verzameling toetsresultaten zitten echter allerlei onregelmatigheden die niet in zo'n rigide model passen. Daarom is het realistischer de scores van de leerlingen te beschrijven met een kansmodel. In paragraaf 3 zal een voorbeeld van zo'n kansmodel gegeven worden.

We zullen hieronder eerst nog wat verder ingaan op de problematiek van het equivaleren van twee examens. We willen examenscores immers niet alleen verklarend beschrijven, maar ook met elkaar vergelijken. In tabel 3 zijn de resultaten op twee examens weergegeven. Om de tabel overzichtelijk te houden is ervan uitgegaan dat het eerste examen bestaat uit de items 1 t/m 4 en dat het tweede examen bestaat uit de items 5 t/m 8. Verder hebben de leerlingen 1 t/m 4 het eerste examen gemaakt en de leerlingen 5 t/m 8 het tweede examen. Op de laatste twee leerlingen gaan we zo dadelijk in. Tenslotte gaan we er even vanuit dat de scores van de leerlingen 0 of 1 zijn.

***Tabel 3**: scores van 10 leerlingen op vragen uit twee meerkeuze-examens*

leer-	examen 1, item:				examen 2, item:				LL
ling	1	2	3	4	5	6	7	8	som
1	0	1	1	1					3
2	0	1	1	1					3
3	0	1	1	0					2
4	1	0	0	0					1
5					1	1	0	1	3
6					1	0	0	1	2
7					1	0	1	0	2
8					1	1	1	0	3
9	1	0			0	1			2
10	0	1			1	1			3
itemsom	2	4	3	2	5	4	2	2	

Het grote probleem bij het equivaleren van examens is dat de invloed van de moeilijkheidsgraad van de examens gescheiden moet worden van de invloed van het niveau van de leerlingen. We mogen er namelijk niet bij voorbaat vanuit gaan dat de score 3 van leerling 1 dezelfde waardering verdient als de score 3 van leerling 5. Hun examens kunnen namelijk van een verschillende moeilijkheidsgraad zijn. Om de scores van de leerlingen te kunnen vergelijken, moeten we in de eerste plaats een verband tussen de examens leggen. In het voorbeeld van tabel 3 is dit verband gelegd door de leerlingen 9 en 10 twee items van het ene en twee items van het andere examen te laten maken. Hierdoor verkrijgen we informatie over de verhouding in moeilijkheidsgraad tussen de examens. Verder moeten we zoeken naar een kansmodel dat:
a) voldoende bij de voorhanden scores past en
b) vergelijkbare schattingen van het vaardigheidsniveau van de leerlingen oplevert.

In paragraaf 3 zal een voorbeeld van zo'n model gegeven worden en van de informatie die zo'n model kan verschaffen over het verband tussen de opgaven uit verschillende examens. In paragraaf 4 worden voorbeelden gegeven van pretest- en posttestprocedures.

3. Het schatten van parameters en scores

De parameters in het Rasch-model
In deze paragraaf zullen we laten zien hoe de metingen van tabel 3 geëquivaleerd kunnen worden aan de hand van een realistisch en nog tamelijk eenvoudig psychometrisch model. Dit model, van de Deense statisticus Georg Rasch, is een kansmodel waarin de kans dat een bepaalde leerling een bepaald item goed beantwoordt, afhankelijk is van de *vaardigheid van de persoon* en van de *moeilijkheid van het item*. In dit model wordt ervan uitgegaan dat de scores van de leerlingen in twee categorieën verdeeld kunnen worden, bijvoorbeeld 'goed' en 'fout', of 1 en 0. Het Rasch-model houdt in dat de kans op een goed antwoord 50% is, indien de vaardigheid van een leerling precies even groot is als de moeilijkheid van het item. Is de vaardigheid kleiner dan de moeilijkheid, dan neemt de kans af, en is de vaardigheid groter dan de moeilijkheid, dan neemt de kans toe.

De procedure voor het equivaleren van examens begint met het schatten van de persoonsparameters θ en itemparameters δ. In de praktijk gebeurt dat met zogenaamde grootste aannemelijkheidsschatters, op de technische details gaan we hier verder niet in. In tabel 4 zijn bij het voorbeeld van tabel 3 de schattingen van de persoons- en itemparameters weergegeven. Dus item 1 heeft een itemparameter $\delta_1 = 1.17$ en leerling 1 heeft een persoonsparameter $\theta_1 = 1.28$.

***Tabel 4**: schattingen van itemparameters δ en persoonsparameters θ van de items en personen uit tabel 3*

	examen 1, item:				examen 2, item:			
	1	2	3	4	5	6	7	8
δ	1,17	-0,03	-0,45	0,61	-0,30	-0,30	0,10	-0,79

	leerlingen									
	1	2	3	4	5	6	7	8	9	10
θ	1,28	1,28	0,32	-0,63	0,55	-0,32	-0,32	0,55	-0,10	1,07

De leerlingen 1 en 5 hadden beiden 3 opgaven goed (tabel 3). Leerling 5 blijkt echter een lagere vaardigheidsparameter te hebben (0.55) dan leerling 1 (1.28). Kennelijk is het examen dat leerling 5 gemaakt heeft gemakkelijker dan het examen dat leerling 1 heeft gedaan.

Het toetsen van het model

De volgende stap is nu dat we moeten nagaan of het model ook acceptabel is. Dit is mogelijk door de feitelijke scores (tabel 3) te vergelijken met de kansen die volgens het model elke leerling heeft op een goede score op elk item. Die kansen kunnen we berekenen met behulp van de nu bekende item- en persoonsparameters. De uitkomsten staan in tabel 5. Als voorbeeld nemen we het antwoord van leerling 4 op item 1. Deze leerling gaf hier een goed antwoord, terwijl de kans daarop onder het model maar erg klein is. Dat pleit dus niet voor het model. In de praktijk wordt er een groot aantal statistische toetsen uitgevoerd die alle gebaseerd zijn op vergelijkingen van geobserveerde en onder het model verwachte waarden. Die toetsen zijn zo geconstrueerd dat allerlei specifieke aspecten van het model geëvalueerd worden. Dat heeft als voordeel dat in het geval een model niet blijkt te passen, we ook weten welke aspecten we verder moeten aanpassen om het wel acceptabel te krijgen. Het model wordt net zolang uitgebreid tot uit de toetsing blijkt dat het acceptabel is.

Het schatten van scores

Op grond van onze schattingen van de vaardigheid van de leerlingen (persoonsparameters) en van de moeilijkheid van de items (itemparameters) kunnen we voor elke leerling voorspellen welke score deze zou behalen op elke willekeurige verzameling van items uit de dataset, ook items die deze leerling zelf niet gemaakt heeft. Volgens het Rasch-model is de kans op een goed antwoord immers alleen afhankelijk van de persoons- en van de itemparameter, en die hebben we uit de overige gegevens kunnen schatten. Dit biedt de mogelijkheid om de gegevens te weten te komen die we nodig hebben voor de equivalering, namelijk hoe goed de kandidaten van een bepaald jaar het examen van een ander jaar gemaakt zouden hebben.

***Tabel 5**: kans op een goed antwoord en verwachte totaalscore op niet gemaakte examens volgens het Rasch-model gegeven de geschatte parameters uit tabel 4*

leer-	examen 1, item:				examen 2, item:				verw.	verw.
ling	1	2	3	4	5	6	7	8	ex 1	ex 2
1	.53	.79	.85	.66	.83	.83	.76	.89		3.31
2	.53	.79	.85	.66	.83	.83	.76	.89		3.31
3	.30	.59	.68	.43	.65	.65	.56	.75		2.61
4	.14	.35	.45	.22	.42	.42	.33	.54		1.71
5	.35	.64	.73	.49	.70	.70	.61	.79	2.21	
6	.18	.43	.53	.28	.50	.50	.40	.61	1.42	
7	.18	.43	.53	.28	.50	.50	.40	.61	1.42	
8	.35	.64	.73	.49	.70	.70	.61	.79	2.21	

In tabel 5 zijn voor alle items de kansen op een goed antwoord vermeld. Per examen is bovendien per leerling de te verwachten totaalscore gegeven voor het examen dat hij *niet* gemaakt heeft. De verwachte score op het examen dat hij *wel* gemaakt heeft is gelijk aan de score die hij op dat examen gerealiseerd heeft. Zo is de kans dat leerling 3 item 5 goed zou maken 0.65. Als we de kansen voor deze leerling optellen over alle items uit examen 2, krijgen we als resultaat de zogenaamde verwachte totaalscore op dit examen: 2.61. De score van deze leerling op het andere examen (examen 1) was 2. Uit de hogere score op examen 2 zien we opnieuw dat dit gemakkelijker was dan examen 1.

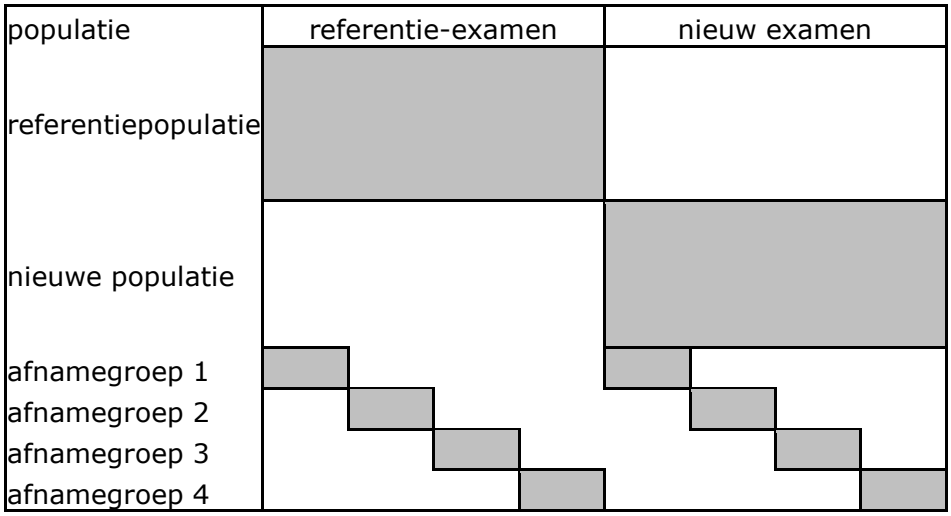
De praktijk

We hebben aan de hand van een eenvoudig voorbeeld laten zien dat het met behulp van een model met persoons- en itemparameters mogelijk is scores te schatten op items die leerlingen zelf niet gemaakt hebben. Deze schattingen maken het vervolgens mogelijk scores op examens te schatten. De praktijk is echter complexer: de modellen waarmee gewerkt moet worden blijken ingewikkelder. Dat is bijvoorbeeld het geval voor examens met open vragen. Daar moet niet alleen de kans berekend worden dat de leerling de vraag geheel goed of geheel fout beantwoord, maar ook welke kans de leerling op elke mogelijke deelscore heeft. Voorts houdt het Rasch-model geen rekening met bepaalde onderlinge verschillen tussen items of leerlingen. Bij de toetsing van het Rasch-model blijkt vaak dat het niet bij de data past. Er moeten dan ingewikkelder modellen gebruikt worden. Ook deze ingewikkelde modellen zijn nog steeds kansmodellen met aparte parameters voor de items en de personen. Alleen is het niet meer zo dat het bij een item- of persoonsparameter nog steeds om één getal voor ieder item en iedere persoon hoeft te gaan. Deze modellen zijn juist ontwikkeld om acceptabele voorspellingen te kunnen doen wanneer items verschillen in discriminerend vermogen of wanneer leerlingen in één deel van de examenstof beter zijn dan in een ander. De procedure is verder zoals hierboven beschreven: we beginnen met een eenvoudig model, schatten de parameters in het model en evalueren de modelpassing. Als het model niet past, wordt het model, gebruikmakend van de uitkomsten van de statistische toetsen, uitgebreid tot een ingewikkelder model. Daarmee wordt doorgegaan tot een passend model gevonden is.

Al deze modellen stellen echter wel bepaalde eisen aan de wijze waarop scores verzameld worden, zoals het verband dat er moet zijn tussen de te vergelijken examens. In paragraaf 4 worden voorbeelden gegeven van de manier waarop dit bereikt kan worden: de pretest en de posttest procedures.

4. Gegevensverzamelingen voor het equivaleren van examens

In deze paragraaf zullen we ingaan op de manier waarop items over leerlingen verdeeld worden bij de gegevensverzameling ten behoeve van de equivalering. We onderscheiden twee methoden, die we zullen aanduiden met de pretest opzet en de posttest opzet. Uitgangspunt vormt bij beide methoden een referentie-examen waarvan we de norm willen overbrengen op het nieuwe examen.



***Figuur 1**: de posttest opzet*

De posttest opzet
Ook in figuur 1 zijn de items van links naar rechts weergegeven en de leerlingen van boven naar beneden. De gearceerde rechthoeken staan voor geobserveerde antwoorden. In deze figuur zijn dus scores op twee examens weergegeven. Behalve de scores van de twee oorspronkelijke populaties, één bestaande uit de kandidaten die het referentie-examen gemaakt hebben en de ander bestaande uit de kandidaten van het lopende examenjaar, zijn ook scores verzameld van vier afnamegroepen. Iedere afnamegroep heeft een stukje van het eerste en een stukje van het tweede examen gemaakt. Deze afnamegroepen dienen om de resultaten van de twee examens met elkaar in verband te brengen: de opgaven die zij maken uit het referentie-examen kunnen immers zowel met de rest van het referentie-examen verbonden worden als met de opgaven uit het nieuwe examen. Deze kunnen verbonden worden met de rest van het nieuwe examen, en zo kunnen dus indirect het referentie-examen en het nieuwe examen met elkaar verbonden worden.

Het vaardigheidsniveau van de afnamegroepen hoeft niet hetzelfde te zijn als die van de examengroepen. Hun scores worden immers in het model weergegeven door aparte persoonsparameters. Wel moeten hun scores in het model passen. Dat is niet het geval als ze op hun toetsjes bijvoorbeeld puur zouden gokken. Het is dus belangrijk dat de leerlingen van de afnamegroepen op hun toetsjes serieus, en gebruikmakend van hun vaardigheid, antwoord geven. Een goede motivatie van de leerlingen is hiervoor essentieel. De statistische toetsen waarmee de passing van een model geëvalueerd wordt, signaleren dat de scores van de afnamegroepen niet passen in het model voor de twee examens. Het is dan verder onmogelijk om op basis van dit model te equivaleren.

In tabel 6 is nog eens aangegeven hoe men zich de equivalering moet voorstellen. De gearceerde gebieden van de datamatrix bevatten de geobserveerde gegevens van de referentiepopulatie, de nieuwe populatie en drie groepen leerlingen die gebruikt zijn om de twee examens met elkaar in verband te brengen. Net als in het voorbeeld van tabel 5 is ook in dit voorbeeld het Rasch-model gebruikt.

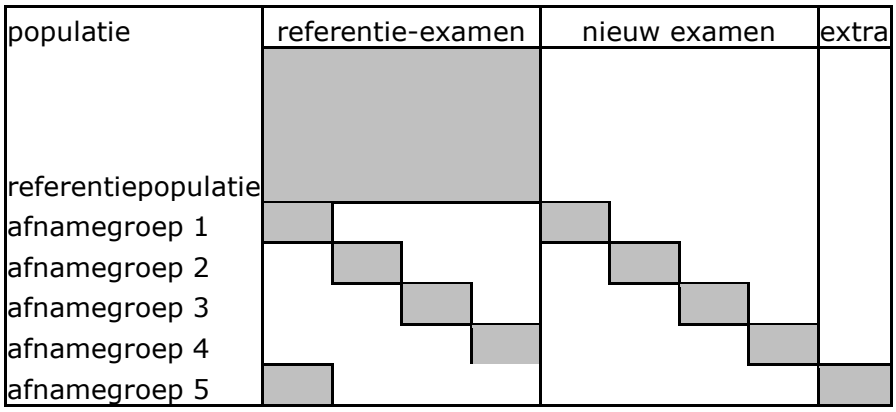
Na het schatten van de itemparameters (δ) en de vaardigheidsparameters van de personen (θ) kan men een schatting maken van de antwoorden van de personen op items die ze niet gemaakt hebben. Deze schattingen staan in de niet-gearceerde delen van de data-matrix in tabel 6. Uit deze schattingen kan men vervolgens de scores schatten die de referentiepopulatie gerealiseerd zou hebben als ze het nieuwe examen gemaakt zouden hebben. Op dezelfde manier kan men ook de scores schatten van de nieuwe populatie als zij het referentie-examen gehad zouden hebben. De gerealiseerde en verwachte scores op het referentie-examen en nieuwe examen staan in laatste twee kolommen van tabel 6. Merk op dat het ook mogelijk is de verwachte scores van de drie experimentele groepen op de twee examens te schatten.

***Tabel 6**: schatting van scores in de posttest opzet*

	referentie-examen						nieuw examen							scores	
	1	2	3	4	5	6	7	8	9	10	11	12		ref	nw
	1	0	0	0	0	0	0,2	0,1	0,1	0,1	0,0	0,0	-2,544	1	0
	1	1	0	0	0	0	0,4	0,3	0,2	0,2	0,1	0,0	1,243	2	1
	1	1	1	0	0	0	0,6	0,6	0,5	0,4	0,2	0,1	-0,221	3	2
ref.	1	1	1	1	0	0	0,8	0,8	0,7	0,6	0,4	0,2	0,742	4	4
pop.	1	1	1	1	1	0	0,9	0,9	0,9	0,8	0,7	0,4	1,773	5	5
	1	0	0	0	0	0	0,2	0,1	0,1	0,1	0,0	0,0	-2,544	1	0
	1	1	0	0	0	0	0,4	0,3	0,2	0,2	0,1	0,0	-1,243	2	1
	0	1	1	0	0	0	0,4	0,3	0,2	0,2	0,1	0,0	-1,243	2	1
	1	0	1	1	0	0	0,6	0,6	0,5	0,4	0,2	0,1	-0,221	3	2
	1	1	0	1	1	1	0,9	0,9	0,9	0,8	0,7	0,4	1,773	5	5
	0,9	0,5	0,4	0,1	0,1	0,0	1	0	0	0	0	0	-1,276	2	1
	0,9	0,7	0,6	0,3	0,2	0,1	1	1	0	0	0	0	-0,449	3	2
	0,9	0,5	0,4	0,1	0,1	0,0	0	0	1	0	0	0	-1,276	2	1
	0,9	0,8	0,7	0,4	0,4	0,2	1	1	0	1	0	0	0,285	4	3
	0,9	0,8	0,7	0,4	0,4	0,2	1	0	1	0	1	0	0,285	4	3
nw.	0,9	0,9	0,9	0,6	0,6	0,3	0	1	1	1	0	1	1,063	4	4
pop.	0,9	0,5	0,4	0,1	0,1	0,0	1	0	0	0	0	0	-1,276	2	1
	0,9	0,5	0,4	0,1	0,1	0,0	0	1	0	0	0	0	-1,276	2	1
	0,9	0,8	0,7	0,4	0,4	0,2	1	1	1	0	0	0	0,285	4	3
	0,9	0,5	0,4	0,1	0,1	0,0	0	0	1	0	0	0	-1,276	2	1
	0,9	0,9	0,9	0,6	0,6	0,3	0	1	1	1	1	0	1,063	4	4
	0,9	0,9	0,9	0,8	0,8	0,5	1	1	0	1	1	1	2,013	5	5
afn.	1	0	0,4	0,1	0,1	0	1	0,2	0,2	0,2	0,1	0	-1,354	2	1
gr 1	1	0	0,1	0,0	0,0	0	0	0,1	0,1	0,1	0	0	-2,595	1	0
afn.	0,9	0,9	1	0	0,5	0,3	0,7	1	1	1	0,4	0,2	0,892	4	4
gr 2	0,9	0,8	1	0	0,3	0,1	0,9	0	0	1	0,2	0,1	-0,020	3	3
afn.	0,9	0,9	0,9	0,9	1	1	0,9	0,9	0,9	0,9	1	0	2,454	5	5
gr 3	0,9	0,9	0,8	0,5	1	0	0,8	0,7	0,7	0,6	0	0	0,593	4	3

De pretest opzet
In figuur 2 is een voorbeeld van een pretest opzet weergegeven. Ook in deze opzet gebruiken we de gegevens van de oorspronkelijke populatie van het referentie-examen en zijn de onderdelen van zowel het referentie-examen als van het nieuwe examen verdeeld over een aantal afnamegroepen. Er zijn twee duidelijke verschillen met de posttestafname.

Het eerste verschil is natuurlijk dat er tijdens de pretest nog geen gegevens beschikbaar zijn van de examenpopulatie waarvoor het nieuwe examen bedoeld is (die gegevens kunnen achteraf wel gebruikt worden om de equivalering te evalueren).



***Figuur 2**: de pretest opzet*

Het tweede verschil bestaat uit de aanwezigheid van extra opgaven, naast het in concept vastgestelde nieuwe examen. Het vervangen van opgaven uit het examen door extra opgaven met een andere moeilijkheid is één van de mogelijke methoden om ervoor te zorgen dat het nieuwe examen even moeilijk wordt als het referentie-examen en genormeerd kan worden met een N-term van 1,0. Zodra het nieuwe examen daadwerkelijk afgenomen is, kan men de scores van de examenpopulatie gebruiken om de nauwkeurigheid van de equivalering te evalueren.

In tabel 7 is aangegeven hoe men zich de pretestprocedure moet voorstellen. De gearceerde gebieden van de datamatrix bevatten de geobserveerde gegevens van de referentiepopulatie en vier groepen leerlingen die gebruikt zijn om het nieuwe examen te pretesten. Merk op dat in de opzet van tabel 7 slechts een deel van de items van het referentie-examen in de pretest is meegenomen. Men kan natuurlijk ook alle items van het referentie-examen in de pretest meenemen. Net als in de voorbeelden van tabel 5 en tabel 6 is ook in dit voorbeeld het Rasch-model gebruikt. Eerst zijn de itemparameters en de vaardigheidsparameters van de personen geschat, voor de methode zij men verwezen naar de vorige paragrafen. Daarna kan men een schatting maken van de antwoorden van de personen op items die ze niet gemaakt hebben. Deze schattingen staan in de niet-gearceerde delen van de data-matrix in tabel 7.
Uit deze schattingen kunnen we vervolgens de scores schatten die de referentiepopulatie gerealiseerd zou hebben als ze een examen gemaakt zouden hebben dat is samengesteld uit de nieuwe opgaven. Alle geschatte scores van de referentiepopulatie op het nieuwe examen leveren een geschatte frequentieverdeling op.

De gerealiseerde en verwachte scores op het referentie-examen en de nieuwe opgaven staan in de laatste twee kolommen van tabel 7.

***Tabel 7**: schatting van scores in de pretest opzet*

	referentie-examen						nieuwe opgaven							
	1	2	3	4	5	6	1	2	3	4	5	6	ref ex	nw ex
ref.	1	0	0	0	0	0	0,2	0,5	0,1	0,0	0,8	0,0	1	2
pop.	1	1	0	0	0	0	0,4	0,8	0,2	0,0	0,9	0,0	2	3
	1	1	1	0	0	0	0,6	0,9	0,4	0,1	0,9	0,0	3	4
	1	1	1	1	0	0	0,8	0,9	0,6	0,3	0,9	0,0	4	5
	1	1	1	1	1	0	0,9	0,9	0,9	0,6	0,9	0,0	5	5
	1	0	0	0	0	0	0,2	0,5	0,1	0,0	0,8	0,0	1	2
	1	1	0	0	0	0	0,4	0,8	0,2	0,0	0,9	0,0	2	3
	1	1	1	0	0	0	0,4	0,8	0,2	0,0	0,9	0,0	2	3
	1	1	1	1	0	0	0,6	0,9	0,4	0,1	0,9	0,0	3	4
	1	1	0	1	1	1	0,9	0,9	0,9	0,6	0,9	0,0	5	5
gr 1	1	0	0,3	0,2	0,0	0,0	1	0	0,1	0,0	0,9	0,7	2	4
	0	1	0,3	0,2	0,0	0,0	0	1	0,1	0,0	0,9	0,7	2	4
	1	0	0,5	0,4	0,1	0,0	1	1	0,3	0,1	0,9	0,8	2	4
	0	1	0,3	0,2	0,0	0,0	0	1	0,1	0,0	0,9	0,7	2	3
gr 2	0,9	0,9	1	1	0,4	0,3	0,9	0,9	1	0	0,9	0,9	4	5
	0,9	0,8	1	0	0,2	0,1	0,7	0,9	0	1	0,9	0,9	3	5
	0,9	0,9	1	1	0,4	0,3	0,9	0,9	1	0	0,9	0,9	6	5
	0,9	0,9	1	1	0,4	0,3	0,9	0,9	1	0	0,9	0,9	6	5
gr 3	0	0	0,2	0,1	0,0	0,0	0,2	0,5	0,1	0,0	1	1	0	3
	1	0	0,1	0,0	0,0	0,0	0,1	0,3	0,0	0,0	0	0	1	0
	0	0	0,2	0,1	0,0	0,0	0,2	0,5	0,1	0,0	1	1	0	3
	1	0	0,2	0,1	0,0	0,0	0,2	0,5	0,1	0,0	1	0	1	2
gr 4	0,8	0,6	1	0	0,1	0,0	0	0,8	0,2	0,1	0,9	1	3	3
	0,6	0,3	0	1	0,0	0,0	0	0,6	0,1	0,0	0,8	0	2	2
	0,9	0,8	1	0	0,2	0,1	1	0,9	0,4	0,1	0,9	1	3	4
	0,8	0,6	0	1	0,1	0,0	0	0,8	0,2	0,1	0,9	1	3	3

Merk op dat in het nieuwe examen van het voorbeeld alle gepreteste items gebruikt zijn. Als er meer nieuwe items ter beschikking staan dan voor het nieuwe examen noodzakelijk is (zoals in figuur 2), kan men trachten de items voor het examen zo te kiezen dat de geschatte frequentieverdeling zo dicht mogelijk bij de op het oude examen gerealiseerde frequentieverdeling komt te liggen. Dit alles natuurlijk binnen de randvoorwaarden die de moeilijkheidsgraad van de itemparameters met zich meebrengt.

Toepassing van de uitkomsten van de afnames

In paragraaf 3 is getoond dat we de totaalscores van leerlingen kunnen schatten op examens die ze niet gemaakt hebben. Zo kunnen we van de leerlingen uit de oorspronkelijke populatie van het referentie-examen de scores schatten die ze zouden halen op het nieuwe examen. Van deze geschatte scores maken we een

cumulatieve frequentieverdeling[1]. En door deze naast hun frequentieverdeling op het referentie-examen te leggen, kunnen we de resultaten op beide examens vergelijken (tabel 8). We zien in tabel 8 dat bij de gegeven cesuur van 24/25 op het referentie-examen 21.0% van de referentiepopulatie een score onder de cesuur heeft gerealiseerd. De laatste kolom van tabel 8 bevat de geschatte frequentietabel van de scores die de kandidaten uit de referentiepopulatie zouden behalen op het nieuwe examen. Het nieuwe examen is equivalent met het referentie-examen indien het een equivalente cesuur heeft (zie pag. 1). Het percentage onvoldoendes in het nieuwe examen is bij een cesuur van 24/25 echter 42.2%, een verschil van 21.2%! Het is duidelijk dat het nieuwe examen veel moeilijker is. Wat we daaraan kunnen doen, hangt ervan af of het een vak uit de posttest of uit de pretest procedure betreft. Bij de posttestprocedure is het examen al afgenomen en kan het verschil in moeilijkheid uitsluitend door een verschil in N-term worden gecompenseerd.

[1] De cumulatieve frequentie is een bij een toetsscore behorend getal dat aangeeft hoeveel kandidaten (of welk percentage) de genoemde toetsscore of een lagere hebben behaald. In deze notitie wordt altijd een procentueel cumulatieve frequentieverdeling weergegeven.

***Tabel 8**: cumulatieve frequentieverdeling van het referentie-examen en het nieuwe examen*

Populatie	Referentie-populatie	
	Ref. Ex	Nw. Ex
Score	Cum. freq	Cum. freq
14	0,9	8,1
15	2,1	12,3
16	2,4	13,5
17	3,9	14,7
18	4,8	**19,8**
19	7,5	22,5
20	9,9	24,3
21	12,3	29,3
22	14,7	31,4
23	17,7	38,0
24	**21,0**	42,2
25	23,7	48,5
26	28,7	54,2
27	33,8	56,9
28	39,2	62,6
29	47,3	69,2
30	53,0	74,3
31	59,3	77,2
32	65,6	80,8
33	71,3	81,7
34	76,6	87,1
35	80,5	91,3
36	87,4	94,3
37	90,4	95,5
38	93,4	96,4
39	94,9	97,0
40	96,7	98,5
41	98,5	98,8
42	99,1	98,8
43	99,4	99,4
44	99,4	99,4
45	99,7	99,4
46	99,7	100,0
47	99,7	100,0
48	100,0	100,0
49	100,0	100,0
50	100,0	100,0

Mogelijke acties in de pretest procedure
De pretest procedure biedt enkele beperkte mogelijkheden om het nieuwe examen zelf aan te passen, zodat de equivalente N-term rond de ,0 kan worden vastgesteld.

Dit zijn:

Vervanging: indien er extra opgaven over dezelfde stof voorhanden zijn, waarvan de moeilijkheid bekend is (zie de pretest opzet, figuur 2, afnamegroep 5), kunnen deze benut worden om de moeilijkheid van het gehele examen te veranderen. Het effect is vooral merkbaar wanneer het verschil in moeilijkheid tussen de opgaven aanzienlijk is.

Weglating: in sommige gevallen kan een vraag worden weggelaten, bijvoorbeeld de laatste vraag in een serie die erg moeilijk is. Weglating van een zogenaamde 'opstapvraag' daarentegen (een makkelijke, eerste vraag in een serie), kan tot gevolg hebben dat de volgende vragen slechter gemaakt worden (dit blijkt niet uit heranalyse van dezelfde gegevens, maar pas na herafname!). De mogelijkheden tot weglating zijn beperkt, omdat dit de dekking van de stof vermindert.

Uiteraard kunnen ook de opgaven zelf redactioneel bewerkt worden of kan een vraag makkelijker worden gemaakt door meer aanwijzingen te geven. Inhoudelijke aanpassingen aan de opgave zelf maken de resultaten van de pretest echter onbruikbaar. Een aangepaste vraag wordt als een nieuwe opgave beschouwd omdat we de eigenschap van die vraag niet kennen.

Mogelijke acties in de posttest procedure
In de posttest procedure is het nieuwe examen afgenomen en beoordeeld. Dit examen kan alleen nog maar door de keuze van de N-term geëquivaleerd worden. We doen dat door eerst op zoek te gaan naar de equivalente cesuur. De cesuur op het nieuwe examen wordt nu zó bepaald, dat hij in de referentiepopulatie zal leiden tot eenzelfde percentage onvoldoendes als op het referentie-examen zelf.
In tabel 8 leidt dit tot de cesuur 18/19, want het percentage onvoldoendes hierbij ligt het dichtste bij het percentage van 21.0:
het referentie-examen met een cesuur van 24/25 is equivalent met het nieuwe examen met als cesuur 18/19.
Aan de hand van de gegevens in tabel 9 valt tenslotte af te lezen wat de gevolgen van deze equivalente cesuur zijn voor de nieuwe populatie: de cesuur van 18/19 leidt tot slechts 15.8% onvoldoendes op het nieuwe examen.

***Tabel 9**: cumulatieve frequentieverdelingen van beide examens voor beide populaties*

Populatie:	Referentie-populatie		Nieuwe populatie	
	Ref Ex	Nw Ex	Nw Ex	Ref Ex
Score	Cum. freq	Cum. freq	Cum. freq	Cum. freq
14	0,9	8,1	2.7	.0
15	2,1	12,3	5.8	.3
16	2,4	13,5	7.3	.3
17	3,9	14,7	10.3	.6
18	4,8	**19,8**	**15.8**	1.5
19	7,5	22,5	19.1	2.1
20	9,9	24,3	27.3	4.5
21	12,3	29,3	34.5	8.2
22	14,7	31,4	39.1	10.6
23	17,7	38,0	44.5	14.2
24	**21,0**	42,2	50.9	20.9
25	23,7	48,5	56.1	28.5
26	28,7	54,2	63.3	34.2
27	33,8	56,9	68.8	41.5
28	39,2	62,6	73.6	47.9
29	47,3	69,2	78.5	53.6
30	53,0	74,3	82.1	58.2
31	59,3	77,2	86.4	69.7
32	65,6	80,8	89.7	77.3
33	71,3	81,7	92.1	82.1
34	76,6	87,1	93.9	84.2
35	80,5	91,3	97.3	88.2
36	87,4	94,3	98.2	92.1
37	90,4	95,5	99.7	94.8
38	93,4	96,4	99.7	98.2
39	94,9	97,0	99.7	98.8
40	96,7	98,5	100.0	99.7
41	98,5	98,8	100.0	99.7
42	99,1	98,8	100.0	99.7
43	99,4	99,4	100.0	99.7
44	99,4	99,4	100.0	100.0
45	99,7	99,4	100.0	100.0
46	99,7	100,0	100.0	100.0
47	99,7	100,0	100.0	100.0
48	100,0	100,0	100.0	100.0
49	100,0	100,0	100.0	100.0
50	100,0	100,0	100.0	100.0