Wat betekent het twee examens aan elkaar te equivaleren?

Op grond van de principes van eerlijkheid en transparantie van toetsing mogen kandidaten verwachten dat het examen waarvoor ze opgaan gelijkwaardig is aan de oudere examens, waarmee ze zich hebben voorbereid. Maar wat betekent gelijkwaardigheid en hoe is die in de praktijk te bereiken? In dit artikel wordt de methodiek van het equivaleren van twee examens nader uitgelegd. Het equivaleren van twee toetsen is het vergelijken van de moeilijkheid van de toetsen.

Gelijkwaardige centrale examens betekenen in het ideale geval identieke examens, dus: identieke opgaven, identieke afnamecondities, identieke beoordeling en identieke waardering. In de praktijk is het onmogelijk om elk jaar identieke examens af te nemen. Daarom voeren we een normhandhavingsprocedure uit waarbij we de volgende definities gebruiken:

Gelijkwaardige examens: elk examen is een steekproef die volgens dezelfde regels wordt getrokken uit de examenstof en de in het examenprogramma onderscheiden vaardigheden.

Gelijkwaardige standaard: de grens tussen voldoende en onvoldoende wordt op gelijkwaardige examens zodanig bepaald dat deze, voor één en dezelfde populatie, op ieder examen tot hetzelfde percentage onvoldoendes leidt.

Equivalente examens: gelijkwaardige examens genormeerd volgens een gelijkwaardige standaard.

Even moeilijk, of niet?
Wanneer twee gelijkwaardige examens niet even moeilijk blijken te zijn, kan men ze toch nog equivalent maken door tegenover de ongelijke moeilijkheid een daartegen opwegende ongelijkheid in de cesuur te stellen. Voor het examen met een hogere moeilijkheid stelt men dan een hogere normeringsterm (en een lagere cesuur) vast. Dat kan natuurlijk alleen indien de omvang van het verschil in moeilijkheid bekend is.

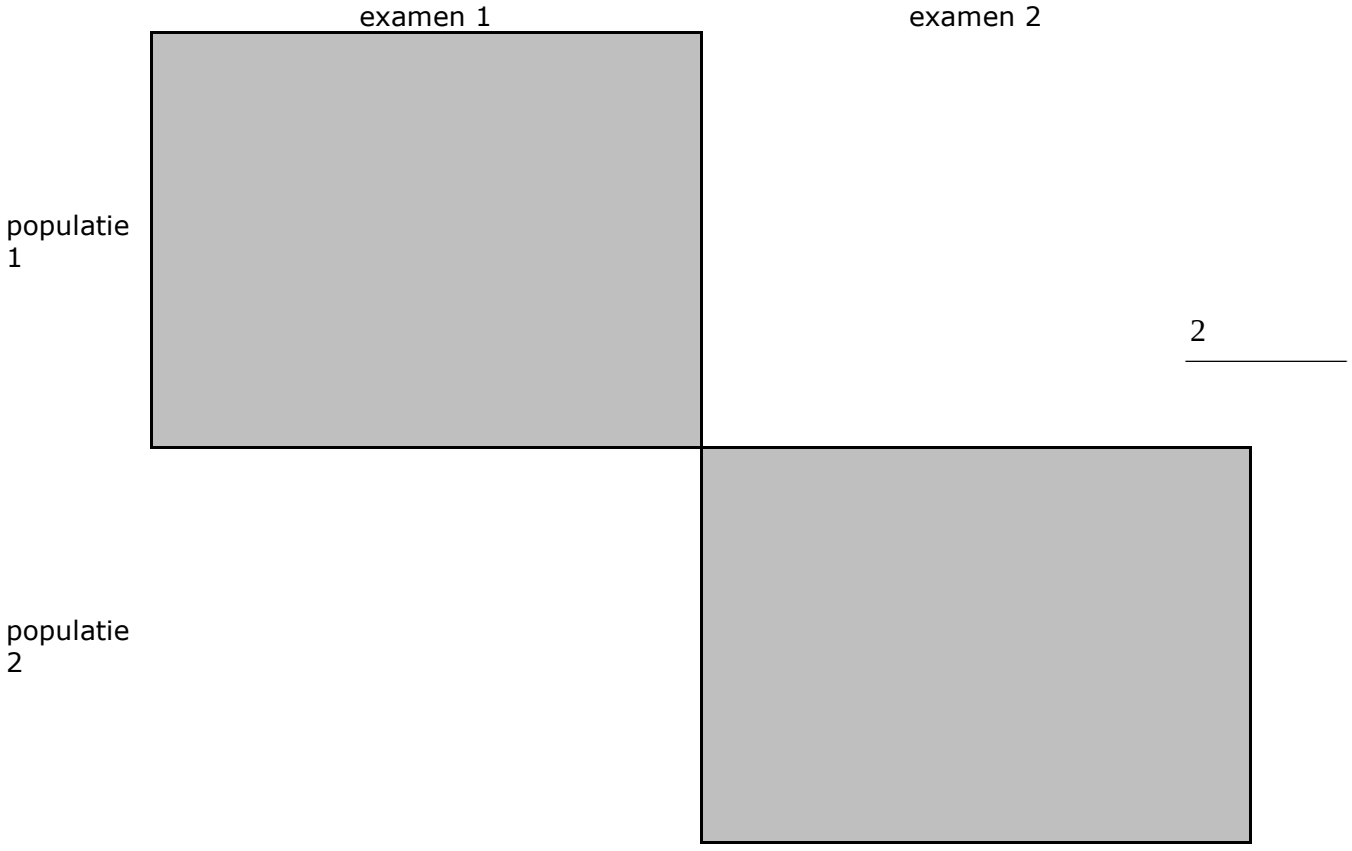
De constructeurs van examens weten dat het moeilijk is om vooraf te schatten hoe moeilijk een examen zal zijn. Maar ook achteraf valt uit de gebruikelijke gegevens over de examenresultaten niet met zekerheid op te maken of een examen moeilijker is dan de vorige jaren, laat staan hoeveel precies. Dat komt omdat in de resultaten, zowel in de gemiddelde score als in het percentage onvoldoendes, twee dingen tegelijk tot uitdrukking komen, namelijk: de moeilijkheid van het examen en de vaardigheid van de populatie (= de groep kandidaten die het examen gemaakt heeft). Wanneer de gemiddelde score lager is dan die van het vorige jaar weet men niet of de opgaven hiervan de oorzaak waren of de populatie, of beide. Dat maakt het – zonder nadere aannamen – niet eenvoudig te bepalen hoeveel punten nodig zijn voor een voldoende. Als het examen moeilijker was geweest dan het vorige jaar, kon dit gecorrigeerd worden door bij minder punten al een voldoende toe te kennen. Maar als het examen even moeilijk was geweest, moest men concluderen dat de kandidaten minder goed waren voorbereid. Was het examen nu even moeilijk of niet?

Equivaleren
Het equivaleren van twee toetsen is het vergelijken van de moeilijkheid van de toetsen. Zo kan berekend worden welke score op de ene toets overeenkomt met een bepaalde score op de andere toets. In algemene zin gaat het bij examens om

meetinstrumenten. Goede meetinstrumenten behoren geijkt te zijn. Bij de meeste meetinstrumenten is er sprake van 'standaard' eenheden. Zo kennen we in de natuurkunde het SI-systeem van de zeven basiseenheden. Het gaat daarbij om fundamentele eenheden voor natuurkundige grootheden. Bij examens kennen we die standaardeenheden niet. Weliswaar wordt er gewerkt met scorepunten, maar het aantal scorepunten dat een kandidaat haalt, is net zo zeer afhankelijk van het examen dat hij maakt als van de mate waarin hij de te meten vaardigheid beheerst. De behaalde scores zijn eigenlijk alleen te vergelijken binnen de groep van kandidaten die dat examen heeft afgelegd.

Wanneer twee verschillende examens door twee verschillende populaties worden afgelegd en we beschikken over de afnamegegevens van de beide examens, dan valt er zonder verdere aannamen te doen weinig te zeggen over eventuele verschillen in moeilijkheidsgraad tussen de examens of eventuele verschillen in vaardigheidsniveau van beide populaties.

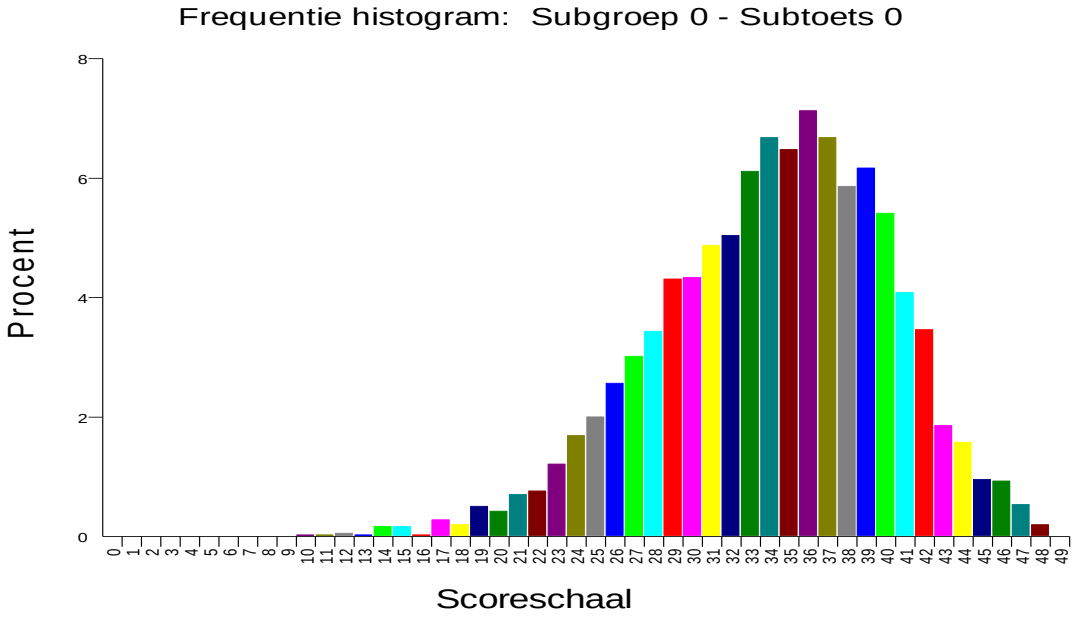
Het afnamedesign ziet er dan als volgt uit:



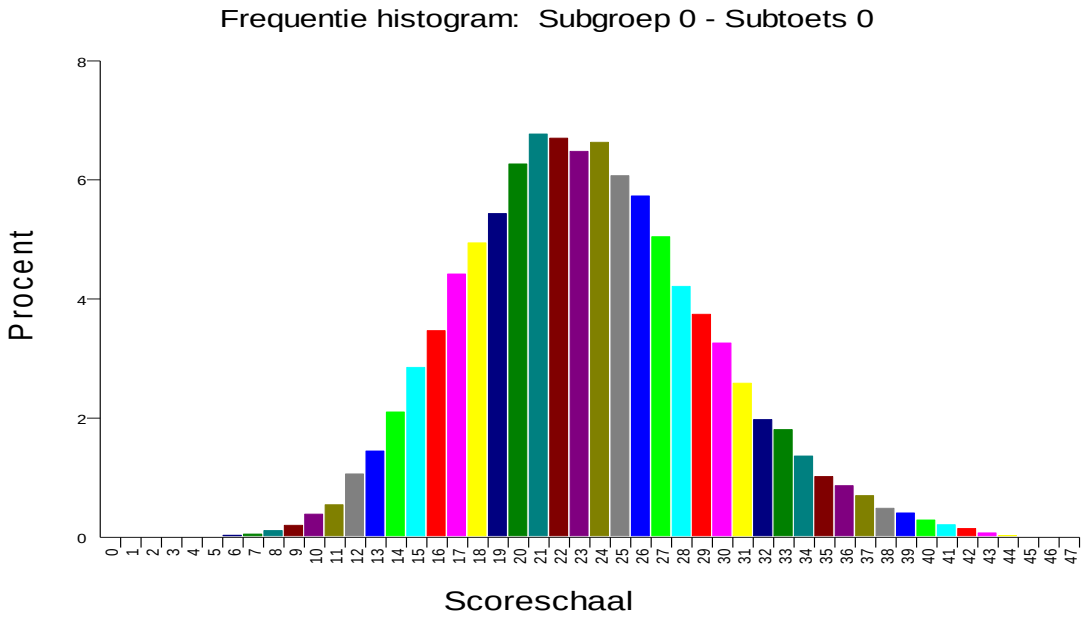
Op basis van deze gegevens zijn er binnen een examen wel uitspraken te doen over wat de moeilijkste opgave van het examen is of welke leerling de hoogste score heeft behaald. De beide examens zijn echter evenmin als de beide populaties goed met elkaar te vergelijken. We kunnen dus geen uitspraak doen in de geest van 'het tweede examen is moeilijker' of 'de populatie die het tweede examen heeft gemaakt, is minder vaardig'. We gebruiken twee verschillende examens en die leveren allebei een verschillende scoreschaal op. Binnen een afname kunnen we de kandidaten ordenen op basis van hun behaalde score, maar kandidaten die verschillende examens hebben gemaakt kunnen we niet op basis van de behaalde score met elkaar vergelijken.

Twee examens, één populatie
Twee examens kunnen wel met elkaar vergeleken worden wanneer ze beide door dezelfde populatie zijn afgelegd. We werken dat idee hieronder verder uit door de resultaten te presenteren van één populatie die twee keer een volledig examen heeft afgelegd. Die examens zijn qua dekking van de eindtermen gelijkwaardig. Beide toetsen zijn samengesteld aan de hand van dezelfde toetsmatrijs en hebben dus een vergelijkbare dekking van de leerstof. Ook golden voor beide examens dezelfde afnamecondities. Er was dus geen verschil in afnameduur en toegestane hulpmiddelen waarover de kandidaten tijdens de afname konden beschikken.

Het eerste examen blijkt een behoorlijk makkelijk examen te zijn geweest. Het tweede examen was zonder meer pittig. Hieronder worden de behaalde scores op de beide examens in een histogram weergegeven. Uit het histogram is meteen te zien dat de scoreverdeling van beide examens qua vorm een redelijke gelijkenis vertoont. Kijken we naar de plaats van het gemiddelde op de schaal dan zien we dat op het eerste examen zowel absoluut als relatief hoger gescoord is.



De kandidaten konden voor dit examen maximaal 49 scorepunten halen. Gemiddeld halen ze daar - met een gemiddelde score van 34,2 - bijna 70% van. Het tweede examen dat deze populatie maakt is een stuk moeilijker. Op het tweede examen kan men maximaal 47 scorepunten halen. Gemiddeld haalt de populatie daar - met een gemiddelde score van slechts 23,3 - nog geen 50% van.



Stel dat het eerste examen het referentie-examen is en voor dat examen een N-term is vastgesteld van 0,0. Dat is een relatief lage N-term, maar bij een makkelijk examen hoort een lage N-term. Bij deze N-term heeft 21,6% van de kandidaten een onvoldoende. De grens tussen voldoende en onvoldoende ligt dan tussen de scores 29 en 30. Bij score 29 heeft een kandidaat een 5,3 en bij score 30 een 5,5.

Het heeft geen zin deze grensscore over te brengen. Stel dat we dat wel zouden doen en dus ook op de tweede toets de grensscore 29/30 zouden hanteren als grens tussen voldoende en onvoldoende. Wat zou dat voor de resultaten op het tweede examen betekenen? Een blik op het histogram laat zien dat dan zeker zo'n driekwart van de kandidaten een onvoldoende zou halen. Als voor hetzelfde vak twee examens worden afgelegd en van eenzelfde groep kandidaten slaagt 78% voor de ene toets en op de andere toets zakt 85%, dan valt moeilijk vol te houden dat die examens equivalent zijn. Het exact overnemen van de grensscore is dus geen goede methode. Het doet geen recht aan verschillen in moeilijkheidsgraad en zelfs als de examens al even moeilijk zouden zijn, houdt het geen rekening met verschillen in schaallengte.

In tabel 1 hebben we de frequentieverdeling van de populatie op beide toetsen in tabelvorm weergegeven. In de eerste kolom staan de cijfers die zijn toegekend aan de scores op toets 1 bij de N-term van 0,0. We zien dat voor het cijfer 5,3 een score van 29 nodig is. We zien verder op die regel dat 869 kandidaten die score hebben gehaald. Dat is 4,3% van de totale groep. We zien dat 4356 kandidaten een score van 29 of lager hebben gehaald en dat betekent dat - bij een N-term van 0,0 - 21,6% van de kandidaten een onvoldoende heeft gehaald. Er moeten dus op dit examen om aan een voldoende te komen tenminste 30 scorepunten behaald worden. Door dit examen met deze N-term aan te wijzen als referentie-examen wordt een prestatie-eis vastgelegd.

We gaan nu deze prestatie-eis overbrengen op het tweede examen. Om dat te doen zoeken we eerst naar de grens tussen voldoende en onvoldoende. Het histogram uit figuur 1 wordt hieronder als een frequentietabel weergegeven (zie tabel 1). Per mogelijke score wordt aangegeven welk cijfer daarbij hoort volgens een N-term van 0,0. Vervolgens wordt aangegeven hoeveel kandidaten een bepaalde score hebben gehaald. In de kolom cumulatieve verdeling wordt

aangegeven hoeveel procent van de kandidaten een bepaalde score of lager heeft gehaald. Uit die tabel kunnen we aflezen dat de grens tussen voldoende en onvoldoende op deze toets met N-term 0,0 ligt tussen de scores 29 en 30. Als we nu op de tweede toets een equivalente grens tussen voldoende en onvoldoende willen bepalen, moeten we zoeken naar de grens die zo dicht mogelijk in de buurt van de 21,6% onvoldoende ligt.

Als we vervolgens daar de bijbehorende cesuur opzoeken en N-term uitrekenen beschikken we over de equivalente N-term voor toets 2. Omdat de beide examens nogal verschillen in moeilijkheidsgraad is het niet vreemd dat de N-termen nogal van elkaar verschillen. Juist door die verschillende N-termen komt deze populatie op beide toetsen tot eenzelfde resultaat in termen van percentage onvoldoende. Meten we deze populatie met de beide geëquivaleerde toetsen, dan komen beide toetsen tot dezelfde uitspraak over deze populatie: zo'n 22% heeft een onvoldoende resultaat behaald.

Tabel 1

		toets 1				toets 2			
cijfer op toets 1	Score	Abs. Freq.	Abs. Perc.	Cum. Freq.	Cum. Perc.	Abs. Freq.	Abs. Perc.	Cum. Freq.	Cum. Perc.
1,4	4	0	0	0	0	1	0	1	0
1,5	5	0	0	0	0	1	0	2	0,01
1,6	6	0	0	0	0	7	0,03	9	0,04
1,6	7	0	0	0	0	11	0,05	20	0,1
1,7	8	0	0	0	0	23	0,11	43	0,21
1,8	9	0	0	0	0	41	0,2	84	0,42
1,9	10	6	0,03	6	0,03	78	0,39	162	0,8
2,0	11	5	0,03	11	0,06	110	0,55	272	1,35
2,2	12	11	0,06	23	0,11	214	1,06	486	2,41
2,4	13	6	0,03	28	0,14	292	1,45	778	3,86
2,6	14	34	0,17	62	0,31	424	2,1	1202	5,96
2,8	15	34	0,17	97	0,48	575	2,85	1777	8,82
2,9	16	6	0,03	102	0,51	699	3,47	2476	12,29
3,1	17	57	0,28	159	0,79	891	4,42	3367	16,71
3,3	18	40	0,2	199	0,99	996	4,94	4363	**21,65**
3,5	19	102	0,51	301	1,49	1095	5,43	5458	27,08
3,7	20	85	0,42	386	1,92	1263	6,27	6721	33,35
3,9	21	142	0,7	528	2,62	1365	6,77	8086	40,12
4,0	22	153	0,76	681	3,38	1351	6,7	9437	46,82
4,2	23	244	1,21	926	4,59	1306	6,48	10743	53,3
4,4	24	341	1,69	1266	6,28	1337	6,63	12080	59,94
4,6	25	403	2	1670	8,28	1224	6,07	13304	66,01
4,8	26	517	2,56	2186	10,85	1155	5,73	14459	71,74
5,0	27	608	3,01	2794	13,86	1017	5,05	15476	76,79
5,1	28	693	3,44	3487	17,3	849	4,21	16325	81
5,3	29	869	4,31	4356	**21,61**	755	3,75	17080	84,75
5,5	30	875	4,34	5230	25,95	657	3,26	17737	88,01
5,7	31	982	4,87	6213	30,83	521	2,59	18258	90,59
5,9	32	1017	5,04	7229	35,87	398	1,97	18656	92,57
6,1	33	1232	6,11	8461	41,98	364	1,81	19020	94,37
6,2	34	1346	6,68	9807	48,66	275	1,36	19295	95,74
6,4	35	1306	6,48	11113	55,14	205	1,02	19500	96,75
6,6	36	1437	7,13	12550	62,27	175	0,87	19675	97,62
6,8	37	1346	6,68	13896	68,95	141	0,7	19816	98,32
7,0	38	1181	5,86	15077	74,81	98	0,49	19914	98,81
7,2	39	1244	6,17	16321	80,98	82	0,41	19996	99,22
7,3	40	1090	5,41	17411	86,39	58	0,29	20054	99,5
7,5	41	823	4,09	18235	90,48	43	0,21	20097	99,72
7,7	42	698	3,47	18933	93,94	30	0,15	20127	99,87
7,9	43	375	1,86	19308	95,8	15	0,07	20142	99,94
8,2	44	318	1,58	19626	97,38	6	0,03	20148	99,97
6,5	45	193	0,96	19819	98,34	3	0,01	20151	99,99
8,9	46	187	0,93	20006	99,27	2	0,01	20153	100
9,3	47	108	0,54	20114	99,8	1	0	20154	100
9,6	48	40	0,2	20154	100				
10,0	49	0	0	20154	100				

Equivalering tot in detail bekeken
We zullen hieronder in detail uitwerken hoe de equivalering in zijn werk gaat. Van belang is het idee dat de beide examens door één populatie zijn gemaakt. Tabel 1 liet zien dat de grens tussen voldoende/onvoldoende lag bij de score 29/30 en dat bij die grens 21,6 % een onvoldoende haalt. Om een equivalente grens op te zoeken in de tweede toets kijken we in kolom 10 van tabel 1 naar een percentage dat zo dicht mogelijk bij 21,6 ligt. Dat blijkt 21,65 te zijn. De grens tussen de scores 29 en 30 op het eerste examen is equivalent met de grens tussen de score 18 en 19 op het tweede examen. De equivalente cesuur op toets 2 is dus 18/19. Bij dit percentage onvoldoende en bij deze cesuur hoort een N-term van 1,9. Deze N-term lezen we af uit de normeringstabel. In tabel 2 is voor toets 2 een normeringstabel opgesteld. Voor elke mogelijke N-term wordt aangegeven welk gemiddeld cijfer de populatie haalt, welk percentage onvoldoende en welke cesuur bij die N-term hoort.

***Tabel 2**: normeringstabel voor toets 2*

N-Term	Gemiddeld cijfer	Perc. onvoldoende	Cesuur
0,0	4,5	81,0	28,5
0,1	4,6	76,8	27,9
0,2	4,7	76,8	27,4
0,3	4,8	71,7	26,9
0,4	4,9	71,7	26,4
0,5	5,0	66,0	25,9
0,6	5,1	66,0	25,3
0,7	5,2	59,9	24,8
0,8	5,3	59,9	24,3
0,9	5,4	53,3	23,8
1,0	5,5	53,3	23,2
1,1	5,6	46,8	22,7
1,2	5,7	46,8	22,2
1,3	5,8	40,1	21,7
1,4	5,9	40,1	21,2
1,5	6,0	33,3	20,6
1,6	6,1	33,3	20,1
1,7	6,2	27,1	19,6
1,8	6,3	27,1	19,1
1,9	6,4	21,6	18,5

Uit deze normeringstabel blijkt dat de equivalente N-term voor het tweede examen dus 1,9 moet zijn. Bij het equivaleren van twee examens zoals hierboven beschreven, worden niet alle scores geëquivaleerd, maar eigenlijk alleen de grens tussen voldoende en onvoldoende. Hieronder zetten we de belangrijkste gegevens van beide examens naast elkaar. Een N-term van 0,0 op het eerste examen is equivalent met een N-term van 1,9 op het tweede examen.

	Referentie examen	Geëquivaleerd examen
aantal kandidaten	21560	21560
max. score	49	47
gemiddelde score	34,18	23,33
p-waarde	69,7	49,6
N-term	0,0	1,9
% onvoldoende	21,6	21,7
gemiddeld cijfer	6,3	6,4

Ondanks dat de beide toetsen flink verschillen in moeilijkheidsgraad wordt door de equivalering het resultaat uitgedrukt in cijferpunten vrij redelijk gelijk getrokken. Het zal nooit helemaal gelijk worden getrokken. De vorm van de scoreverdelingen is immers niet volledig identiek en het blijft altijd een benadering. Uit de mogelijke N-termen wordt de meest equivalente gekozen. Hieronder worden de verdelingen tussen voldoende en onvoldoende voor de beide examens grafisch weergegeven nadat de scores in cijfers zijn omgezet.