Toetsing en toetsanalyse

D. N. M. de Gruijter

herziene versie november 2008

VOORWOORD

Deze tekst is met name bedoeld voor geïnteresseerde docenten die zich met de constructie en verwerking van tentamens bezighouden. De tekst behandelt een aantal thema's uit de toetsanalyse waar ik regelmatig met docenten over heb gepraat. Met name het gebruik van item-indices en de bepaling van de zak/slaaggrens zijn regelmatig terugkomende gespreksthema's. Over de bepaling van de zak/slaaggrens is altijd veel te doen geweest. Het is ondoenlijk om alle ooit voorgestelde methoden voor de bepaling van de zak/slaaggrens aan de orde te stellen. Ik heb dan ook voor de behandeling van deze thematiek een selectie gemaakt van voorstellen. Deze selectie wijkt enigszins af van de keuzes die andere auteurs hebben gemaakt.

Een thema dat zonder problemen kan worden overgeslagen is *equivaleren*. Aan de voorwaarden voor het formeel equivaleren van verschillende toetsen wordt bij door docenten geconstrueerde toetsen meestal niet voldaan (zoals: hoge toetskwaliteit, redelijk grote groepen studenten, psychometrische expertise). De reden om het onderwerp toch aan te snijden is dat docenten die de Nederlandse literatuur over toetsing lezen, het onderwerp kunnen tegenkomen.

Inhoudsopgave

1. Inleiding

Onderwijs en toetsing zijn onverbrekelijk met elkaar verbonden. Toetsing kan plaats vinden aan het eind van het onderwijs. Dan wordt nagegaan of de studenten zich in voldoende mate de beoogde kennis en vaardigheden hebben eigen gemaakt. Studenten bij wie dat niet het geval is, zakken voor het desbetreffende tentamen. De toetsing wordt summatief gebruikt.

Ook tijdens het onderwijs kan worden getoetst. Daarbij kan worden nagegaan in hoeverre de studenten op de goede weg zijn. Informatie uit de toets kan worden gebruikt om de studenten over sterke en zwakke punten te informeren. De docent krijgt uit de toetsresultaten ook informatie over de effectiviteit van het onderwijs tot dan toe. In deze context wordt de toets formatief gebruikt.

Een toets kan zowel een formatief als een summatief aspect hebben.

Bij sommige onderwijsonderdelen is er wel sprake van toetsing, maar is de toetsing niet scherp afgebakend van de ondersteuning van de student door de docent. Dat is bijvoorbeeld dikwijls het geval bij scripties en werkstukken waarbij de docent commentaar geeft totdat het eindproduct aan de gestelde eisen voldoet.

De summatieve toetsing moet aan hoge eisen voldoen. De studenten moeten weten wat zij op het tentamen mogen verwachten. Zij kunnen zich dan op de juiste wijze op het tentamen voorbereiden. Het tentamen moet de kennis en vaardigheden meten die in het onderdeel worden nagestreefd: de toets moet valide zijn (Messick, 1989). De toets moet voldoende vragen bevatten en representatief zijn zodat een redelijk accuraat beeld van het prestatieniveau van de studenten kan worden verkregen. De bepaling van de zak/slaaggrens moet verantwoord en helder zijn. Voor suggesties m.b.t. het construeren van items of vragen kan men terecht bij Dousma, Horsten & Brants (1997) en Van Berkel (1999). Voor eisen t.a.v. de toetsing, zie De Groot en Van Naerssen (1969) en De Gruijter (1994).

Toetsen worden afgenomen binnen het kader van een tentamen- en examenrooster. Als een student zakt, kan hij of zij aan een herkansing deelnemen. Een student die ten onrechte zakt, heeft dus meer dan één kans. Daartegenover staat het risico dat een student ten onrechte na één of meer pogingen slaagt (Millman, 1989). Het tentamenrooster kan bovendien de studie-inzet van studenten beïnvloeden (Cohen-Schotanus, 1994); Vos (1992) heeft richtlijnen voor het maken van een rooster gegeven. In de propedeuse kan een opleiding ervoor kiezen om het aantal herkansingen beperkt te houden en het risico van onterechte zakbeslissingen te verkleinen door een lichte vorm van compensatie in de examenregeling op te nemen.

Een analyse van de toetsresultaten is om een aantal redenen van belang. De analyse geeft informatie over de kwaliteit van de toets en mogelijke problemen. Uit een itemanalyse kan blijken dat enkele items niet aan de verwachtingen voldoen. Een meerkeuze-item kan bijvoorbeeld te moeilijk blijken. Achteraf blijkt dat één van de foute alternatieven tot misverstanden heeft kunnen leiden. Op basis hiervan kan de becijfering worden aangepast. Het item kan voor eventueel toekomstig gebruik worden herschreven.

Wij gaan hier in op de resultaten die uit een toetsanalyse kunnen worden verkregen. Veel van de te bespreken gegevens kan de docent verkrijgen als hij of zij gebruik maakt van een standaardpakket voor de verwerking van tentamens. Sommige gegevens kunnen ook eenvoudig worden verkregen via een spreadsheet. Wij gaan ook in op verschillende methoden voor de bepaling van de cesuur voldoende/onvoldoende. Veel methoden maken gebruik van gegevens uit een toetsanalyse. Gegevens uit de toetsanalyse kunnen ook gebruikt worden voor een evaluatie van de gehanteerde cesuur.

2. Verwerking van de toets en rapportage m.b.t. de totaalscores

2.1. Inleiding

Op een gegeven ogenblik doet een student een tentamen of tentamenonderdeel. Hij kan dat op hetzelfde moment doen als andere studenten of alleen. Dat laatste is mogelijk bij bijvoorbeeld een mondeling tentamen of een met de computer afgenomen toets.

Een groepstentamen bevat meestal dezelfde vragen voor alle studenten uit de groep deelnemers. Als studenten op verschillende tijdstippen tentamen doen, ligt het voor de hand dat ieder een andere selectie van vragen krijgt. De docent wil immers niet het risico lopen dat studenten die later tentamen doen, informatie over de inhoud van de toets krijgen en daardoor een niet beoogd voordeel behalen.

Als een toets geheel uit meerkeuzevragen bestaat, wordt dikwijls met goed/fout scoring gewerkt. Voor elke goed beantwoorde vraag krijgt de student 1 punt. Een niet beantwoorde vraag of een fout beantwoorde vraag levert 0 punten op. Bij toetsen die geheel of gedeeltelijk uit open vragen bestaat, worden de antwoorden op een vraag dikwijls op een beoordelingsschaal met meer dan twee waarden beoordeeld. Bij toetsen met open vragen krijgt men bovendien met beoordelaareffecten te maken.

Voor de eenvoud beginnen wij met de analyse van een toets waarbij goed/fout scoring wordt gebruikt. Wij zullen een meerkeuzetoets als uitgangspunt nemen. Het eerste wat wij na het tentamen doen, is het berekenen van de totaalscores van de studenten.

2.2. de gegevensrechthoek

In Tabel 1 staan de antwoorden van 10 studenten op 5 meerkeuzevragen. Voor het berekenen van de totaalscore bij een meerkeuzetoets moeten wij de antwoorden hercoderen naar goed/fout, ofwel 1(voor *goed*) en 0 (voor *fout*). De goede antwoorden van de vragen staan in de zogenaamde sleutel: de sleutel geeft aan welk antwoord bij een vraag goed is. Bij een goed antwoord zetten wij de score 1, bij een fout antwoord of een overgeslagen vraag de score 0. De scores van de vragen uit Tabel 1 staan in Tabel 2.

Tabel 1. De scores van 10 studenten op 5 vragen

studentnummer	Vraag 1	Vraag 2	Vraag 3	Vraag 4	Vraag 5
0 (sleutel)	a	a	c	d	b
1	a	a	a	d	b
2	-	b	b	d	b
3	c	a	b	a	b
4	a	a	c	d	b
5	a	a	b	c	c
6	a	a	c	d	b
7	c	d	c	d	b
8	b	a	c	d	b
9	a	a	c	d	b
10	a	d	c	d	b

De totaalscore is nu gelijk aan de som van de scores op de afzonderlijke items. De totaalscore moet nog in een cijfer worden omgezet. Voordat wij dat doen, bekijken wij de gegevens over het tentamen wat nauwkeuriger. Wij doen een toetsanalyse. De basis daarvoor is de rechthoek met per combinatie van student en vraag een score. Voor 5 vragen en 10 studenten zou de rechthoek eruit kunnen zien zoals de rechthoek in Tabel 2. Wij kunnen de rechthoek in een spreadsheet opzetten (maar bij meerkeuzetoetsen wordt veelal gebruik gemaakt van speciale toetssoftware).

Tabel 2. De scores van 10 studenten op 5 vragen

studentnummer	Vraag 1	Vraag 2	Vraag 3	Vraag 4	Vraag 5	Totaal
1	1	1	0	1	1	4
2	0	0	0	1	1	2
3	0	1	0	0	1	2
4	1	1	1	1	1	5
5	1	1	0	0	0	2
6	1	1	1	1	1	5
7	0	0	1	1	1	3
8	0	1	1	1	1	4
9	1	1	1	1	1	5
10	1	0	1	1	1	4
						gem.=3.60
p-waarde	0.60	0.70	0.60	0.80	0.90	0.72

In de kolom aan de rechterhand staan de totaalscores van de studenten. Student nr. 1 heeft 4 van de 5 vragen goed beantwoord. Wij kunnen de scores ook per vraag optellen. De som voor vraag 1 is gelijk aan 6. In de praktijk berekenen wij echter onmiddellijk de gemiddelde score per vraag: de proportie goede antwoorden of *p*-waarde. Voor vraag 1 is de *p*-waarde gelijk aan 0.60[1].

Het gemiddelde van de 10 totaalscores is gelijk aan 3.60. (voor statistische begrippen, zie Bijlage I). de gemiddelde *p*-waarde is 0.72. Het gemiddelde van de totaalscores is gelijk aan het aantal vragen maal de gemiddelde *p*-waarde.

2.3. De scoreverdeling

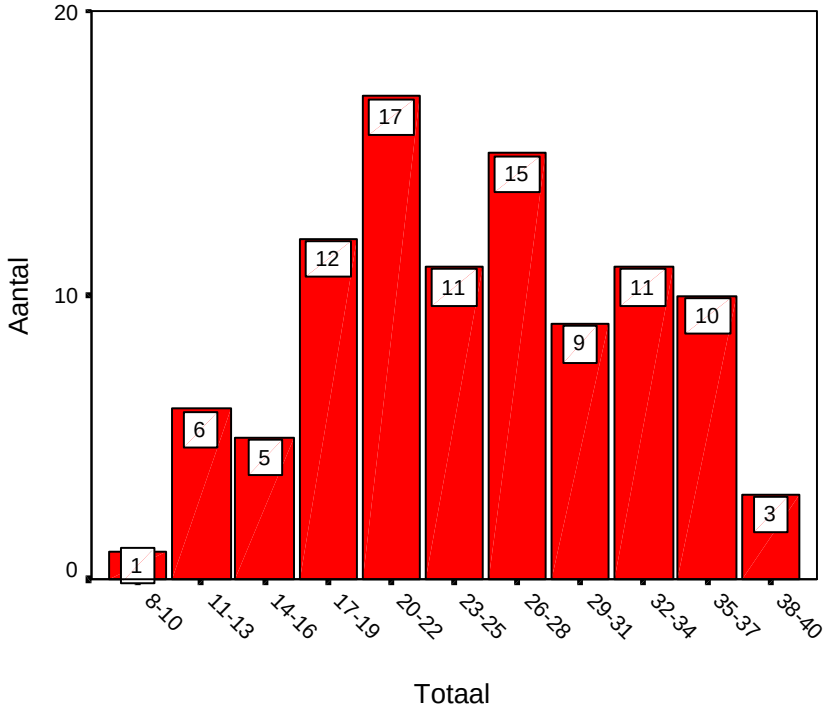
Eerst kijken wij naar de verdeling van de totaalscores (zie Bijlage I voor de uitleg van enkele statistische begrippen). Wij nemen nu een realistischer voorbeeld: een toets bestaande uit 40 vragen, gemaakt door 100 studenten. De scoreverdeling zou eruit kunnen zien als in Tabel 3. Per voorkomende totaalscore staat het aantal

[1] In plaats van een decimale komma, wordt in deze tekst overal een decimale punt gebruikt.

studenten dat deze score heeft behaald en het percentage. Bovendien is het cumulatief percentage in de tabel opgenomen. Zo kunnen wij aflezen dat 48 procent van de studenten een score van 24 of minder heeft behaald. De gemiddelde score is 25.20, de standaardaardafwijking 7.34. De verdeling is grafisch weergegeven in Figuur 1. Voor de overzichtelijkheid zijn de scores in de figuur gegroepeerd in enkele even grote klassen gegroepeerd. De indeling in klassen kan ook anders plaatsvinden. Zodra de cijfers bekend zijn, ligt het voor de hand een frequentieverdeling van (eventueel afgeronde) cijfers te maken.

Tabel 3. De verdeling van de totaalscores

Score	Aantal	Percentage	Cumulatief percentage
9	1	1.00	1.00
10	0	0.00	1.00
11	2	2.00	3.00
12	1	1.00	4.00
13	3	3.00	7.00
14	2	2.00	9.00
15	1	1.00	10.00
16	2	2.00	12.00
17	4	4.00	16.00
18	1	1.00	17.00
19	7	7.00	24.00
20	3	3.00	27.00
21	7	7.00	34.00
22	7	7.00	41.00
23	3	3.00	44.00
24	4	4.00	48.00
25	4	4.00	52.00
26	6	6.00	58.00
27	4	4.00	62.00
28	5	5.00	67.00
29	2	2.00	69.00
30	1	1.00	70.00
31	6	6.00	76.00
32	3	3.00	79.00
33	5	5.00	84.00
34	3	3.00	87.00
35	5	5.00	92.00
36	2	2.00	94.00
37	3	3.00	97.00
38	2	2.00	99.00
39	1	1.00	100.00
Totaal	100	100.00	
gemiddelde:	25.20		
standaardafwijking:	7.34		



Figuur 1. De verdeling van de totaalscores

3. Itemindices

3.1. p-waarde

Eén itemindex zijn wij al tegengekomen: de *p*-waarde. De *p*-waarde geeft de moeilijkheidsgraad van de vraag aan. Eigenlijk is de *p*-waarde een gemakkelijkheidindex: de index is immers de proportie goede antwoorden. Voor de interpretatie van de *p*-waarde moeten wij rekening houden met de aard van de groep studenten die de vraag heeft beantwoord: de *p*-waarde is groepsafhankelijk. De *p*-waarde van een item valt in een groep met herkansers in het algemeen lager uit dan in de groep die voor het eerst tentamen doet: de herkansers hebben dikwijls als groep een lager niveau. Ook andere effecten kunnen van invloed zijn op de hoogte van de *p*-waarden (zie bijv. Van den Brink, 1977). Als de groep die de vraag heeft beantwoord, klein is moeten wij extra voorzichtig zijn: wij moeten met toevalsfluctuaties rekening houden (zie Bijlage I). Een groep van 20 studenten is klein en de gevonden *p*-waarde is minder informatief dan een *p*-waarde bij een groep van meer dan 100 studenten. Als de laatste vragen in het tentamen slecht zijn gemaakt moeten wij alert zijn op de mogelijkheid dat de voor het maken van het tentamen beschikbare tijd te kort was. Ten slotte, de verwachte waarde van de *p*-waarde hangt af van het aantal afleiders. Bij een waar-onwaar vraag zijn er maar twee keuzemogelijkheden. Een student die niets weet, heeft een kans van 50 procent om de vraag door gokken goed te beantwoorden. Bij een vierkeuzevraag is de gokkans ten opzichte van de waar-onwaar vraag gehalveerd.

Een hoge p-waarde kan erop wijzen dat het goede antwoord op de vraag is 'weggegeven'. Een lage *p*-waarde duidt op een te hoge moeilijkheid. De hoge moeilijkheid kan veroorzaakt zijn door een constructiefout, maar het probleem kan ook liggen aan een tekort in het geboden onderwijs.

3.2. r_{it} en r_{ir}

De item-totaalcorrelatie r_{it} en/of de item-restcorrelatie r_{ir} is de tweede belangrijke itemindex. Wij gaan ervan uit dat de betere studenten een vraag relatief vaker goed beantwoorden dan de minder goede studenten: het item discrimineert tussen goede en minder goede studenten. Als dat niet zo is, dan is er iets vreemd aan de hand: de vraag ordent de studenten dan op de verkeerde manier. Als benadering voor hoe goed een student is, wordt de totaalscore op de toets genomen.[2] Wij verwachten dat de correlatie tussen item en totaalscore duidelijk positief is, zeg 0.20 of hoger. De r_{it} heeft een nadeel. In de totaalscore waarmee het item wordt gecorreleerd, komt het item ook voor. De correlatie is dus geflatteerd. Daarom wordt naast of ter vervanging van de r_{it} de item-restcorrelatie gebruikt. Deze correlatie is lager dan de r_{it}, maar moet eveneens positief zijn. Ook bij de correlatiematen moet men met steekproeffluctuaties rekening houden. Bij kleinere groepen studenten kan men uiteraard minder staat maken op de gevonden waarde dan bij grotere groepen. In een later hoofdstuk zal worden ingegaan op de mogelijkheid om de r_{it} en r_{ir} zelf in een spreadsheet te berekenen.

De *p*-waarde en r_{ir} voor de gegevens van onze toets staan in Tabel 4, met de sleutel, de goede alternatieven.

---

[2] Als de toets uit duidelijk te onderscheiden deeltoetsen bestaat, kan een aparte analyse per deeltoets verhelderend zijn.

Tabel 4. De sleutel, de *p*-waarde en r_{ir}

Vraagnummer	goede alternatief	*p*-waarde	r_{ir}
1	1	0.90	0.28
2	1	0.89	0.26
3	4	0.91	0.25
4	2	0.90	0.27
5	3	0.82	0.41
6	2	0.84	0.36
7	1	0.90	0.12
8	4	0.75	0.35
9	2	0.73	0.42
10	3	0.70	0.44
11	2	0.66	0.45
12	1	0.61	0.32
13	3	0.54	0.29
14	3	0.60	0.42
15	2	0.63	0.36
16	4	0.68	0.41
17	4	0.52	0.39
18	1	0.48	0.32
19	1	0.52	0.41
20	1	0.53	0.46
21	2	0.75	0.49
22	2	0.72	0.41
23	4	0.74	0.50
24	1	0.73	0.48
25	3	0.59	0.46
26	3	0.61	0.47
27	3	0.62	0.31
28	4	0.60	0.44
29	4	0.63	0.36
30	1	0.64	0.44
31	2	0.55	0.42
32	4	0.60	0.21
33	3	0.45	0.37
34	3	0.46	0.29
35	1	0.47	0.18
36	4	0.48	0.16
37	4	0.35	0.15
38	2	0.29	0.09
39	2	0.37	0.27
40	3	0.40	0.43

3.3. a-waarde

Het is ook van belang te weten hoe de antwoorden over de foute alternatieven of afleiders verdeeld zijn. Als de *p*-waarde laag is, willen wij weten welke afleider met name wordt gekozen en welke afleiders weinig studenten trekken en dus niet effectieve afleiders zijn. Daarvoor berekenen wij de *a*-waarden, de proporties studenten die een bepaald alternatief hebben gekozen. De *a*-waarden van de

afleiders en het goede alternatief staan voor onze toets in Tabel 5. De *a*-waarde van het goede alternatief (de *p*-waarde) is onderstreept. Uit de gegevens blijkt dat alle vragen (minstens) vier alternatieven kennen: de *a*-waarde is groter dan 0 voor alle alternatieven 4. Er komt één extra kolom in de tabel voor, de kolom waarboven de '0' staat. In deze kolom is de proportie 'niet beantwoord' aangegeven. Zo kunnen wij nagaan of sommige studenten meer tijd voor de beantwoording van de vragen nodig hadden.

Tabel 5. De *a*-waarden van de alternatieven

	a-waarde alternatief				
Vraagnummer	0	1	2	3	4
1		<u>0.90</u>	0.09	0.01	0.0
2		<u>0.89</u>	0.04	0.03	0.04
3		0.01	0.05	0.03	<u>0.91</u>
4		0.03	<u>0.90</u>	0.02	0.05
5		0.03	0.07	<u>0.82</u>	0.08
6		0.04	<u>0.84</u>	0.07	0.05
7		<u>0.90</u>	0.0	0.04	0.06
8		0.08	0.80	0.09	<u>0.75</u>
9		0.10	<u>0.73</u>	0.07	0.10
10		0.09	0.11	<u>0.70</u>	0.10
11		0.10	<u>0.66</u>	0.14	0.10
12		<u>0.61</u>	0.14	0.08	0.17
13		0.17	0.19	<u>0.54</u>	0.10
14		0.08	0.15	<u>0.60</u>	0.17
15		0.10	<u>0.63</u>	0.10	0.17
16		0.12	0.09	0.11	<u>0.68</u>
17		0.15	0.17	0.16	<u>0.52</u>
18		<u>0.48</u>	0.18	0.14	0.20
19		<u>0.52</u>	0.16	0.12	0.20
20		<u>0.53</u>	0.14	0.17	0.16
21		0.10	<u>0.75</u>	0.09	0.06
22		0.11	<u>0.72</u>	0.10	0.07
23		0.08	0.12	0.06	<u>0.74</u>
24		<u>0.73</u>	0.09	0.09	0.09
25		0.17	0.13	<u>0.59</u>	0.11
26		0.16	0.11	<u>0.61</u>	0.12
27		0.18	0.13	<u>0.62</u>	0.07
28		0.13	0.14	0.13	<u>0.60</u>
29		0.15	0.09	0.13	<u>0.63</u>
30		<u>0.64</u>	0.11	0.09	0.16
31		0.16	<u>0.55</u>	0.09	0.20
32		0.14	0.12	0.14	<u>0.60</u>
33		0.22	0.20	<u>0.45</u>	0.13
34		0.18	0.17	<u>0.46</u>	0.19
35		<u>0.47</u>	0.14	0.20	0.19
36	0.01	0.23	0.13	0.15	<u>0.48</u>
37		0.20	0.21	0.24	<u>0.35</u>
38		0.27	<u>0.29</u>	0.21	0.23
39	0.01	0.14	<u>0.37</u>	0.26	0.22
40	0.02	0.19	0.17	<u>0.40</u>	0.22

3.4. M'_{ia} en V'_{ia}

Als de r_{ir} van een vraag negatief is dan weten wij dat de gemiddeld iets betere studenten voor een fout alternatief gekozen hebben. Het is goed om te weten welke afleider voor de beter studenten attractief was. Dat is niet te achterhalen uit tabel 5. Tabel 5 laat alleen zien welke alternatieven aantrekkelijk zijn, maar niet voor welke studenten. Veel toetssoftware geeft de informatie die wel nodig is. Wij introduceren hier enkele indices, die wij hier aangeven met de notatie $M_{ia,}$ M'_{ia} en V'_{ia}.

In de meest eenvoudige vorm wordt gebruikt:

M_{ia} = de gemiddelde totaalscore van alle studenten die bij item *i* alternatief *a* hebben gekozen.

De waarde van M_{ia} kan worden vergeleken met de waarde van *M*, de gemiddelde totaalscore van alle studenten.

Als M_{ia} groter is dan *M*, dan weten wij dat alternatief *a* relatief aantrekkelijk voor betere studenten is. Voor M_{ia} geldt hetzelfde nadeel als voor de item-totaalcorrelatie: bij het goede alternatief kan men gemakkelijk een hogere M_{ia} verkrijgen dan voor de afleiders: bij de berekening van M_{ia} voor het goede alternatief wordt 1 punt voor het goede antwoord op de vraag in de totaalscore meegeteld. Een gecorrigeerd gemiddelde is dus beter:

M'_{ia} = M_{ia} als *a* een afleider is

$M_{ia} - 1$ als a het goede alternatief is.

Wij kunnen ook een verschilscore nemen, bijvoorbeeld:

$$V'_{ia} = M'_{ia} - (M - p_i).$$

Als V'_{ia} voor het goede alternatief negatief is, dan is de item-restcorrelatie ook negatief, als de waarde groter is dan 0, dan is ook de waarde van de item-restcorrelatie groter dan 0. Eventueel kan V'_{ia} nog door de standaardafwijking van de restscores worden gedeeld, om een soort gestandaardiseerde index te verkrijgen. Soms wordt in plaats van een index als V'_{ia} de correlatie tussen het al dan niet kiezen van een afleider en de score op de resttoets gebruikt.

3.5 Het gebruik van de indices

In principe geven de genoemde indices alle informatie die nodig is bij de beoordeling van vragen. Een vraag kan op basis van de waarde van één of meer itemindices als een mogelijk problematische vraag worden aangemerkt. Sommige softwarepakketten geven een waarschuwing indien dat het geval is. Men kan de aandacht het beste eerst richten op de vragen met de slechtste waarden voor de itemindices, vragen met duidelijk negatieve r_{ir} en/of zeer lage *p*-waarden. De itemgegevens hebben des te meer gewicht naarmate het aantal studenten dat de

vraag heeft beantwoord groter is, en met name als de groep studenten voor het grootste deel bestaat uit personen die voor het eerst tentamen doen.

Het is mogelijk dat men bij nadere beschouwing een constructiefout in de vraag ontdekt. In dat geval zal de vraag voor eventueel toekomstige gebruik herschreven moeten worden. Soms blijkt een afleider bij nader inzien toch niet zo 'fout' te zijn. Dan kan de totaalscore opnieuw worden berekend, waarbij dit alternatief ook goed wordt gerekend. Of een vraag blijkt zo slecht te zijn dat het beter lijkt de vraag te laten vallen. Een dergelijk besluit moet overigens niet te snel genomen worden: studenten die de vraag wel goed hebben beantwoord, zouden zich door deze maatregel gedupeerd kunnen voelen.

4. Betrouwbaarheid

De betrouwbaarheid van een toets is de mate waarin de geobserveerde scoreverschillen ware scoreverschillen reflecteren. De betrouwbaarheid wordt gedefinieerd als de verhouding van de variantie van de ware scores en de variantie van de geobserveerde scores (zie Bijlage II). De betrouwbaarheid is een getal tussen de 0 en 1.

De betrouwbaarheid is zowel afhankelijk van de toets zelf als van de groep personen waarbij de toets is afgenomen. In de teller van de betrouwbaarheid staat de variantie van de ware scores en bij een homogene populatie is die kleiner dan bij een heterogene populatie.

De betrouwbaarheid van een toets kan worden beschouwd als de correlatie tussen een test en een tweede test met precies dezelfde eigenschappen, een zogenaamde parallelle test. Wij leggen uit. Veronderstel dat wij over een test en een parallelle test beschikken. In Tabel 6 is de samenhang tussen een test en een tweede, parallelle test weergegeven. In de tabel zijn de scores van 1000 studenten voor de twee tests tegen elkaar afgezet. De correlatie tussen de tests bedraagt 0.80. De betrouwbaarheid van beide tests is 0.80.

De betrouwbaarheid van een toets kan worden geschat met behulp van een betrouwbaarheids-coëfficiënt. Een veel gebruikte coëfficiënt is coëfficiënt alpha (zie Bijlage II). In de literatuur vindt men ook de naam KR20, coëfficiënt alpha voor items met 0-1 scores. Coëfficiënt alpha is een (onder)schatting van de betrouwbaarheid. Een hogere schatting van de betrouwbaarheid is Guttman's λ_2 (De Gruijter & Van der Kamp, 2008).

Laten wij teruggaan naar het voorbeeldtentamen dat wij tot nu toe hebben gebruikt. De waarde van coëfficiënt alpha voor onze 40-item toets bedraagt 0.87.

Waar gebruiken wij de betrouwbaarheid voor? Er zijn verschillende toepassingsmogelijkheden.

Eén toepassing is het bepalen van de lengte van een toets. In veel toepassingen is een hoge betrouwbaarheid vereist. In die toepassingen gaat het er om dat wij nauwkeurig onderscheid tussen personen uit een populatie kunnen maken. Als de betrouwbaarheid laag uitvalt, dan weten wij dat wij meer vragen moeten gebruiken. Bij het toetsen van onderwijsprestaties is een hoge betrouwbaarheid niet zonder meer vereist. Daar gaat het er om dat wij met een redelijke mate van nauwkeurigheid kunnen nagaan of een student aan de eisen voldoet, niet om een onderscheid tussen studenten onderling (zie Hambleton & Novick, 1973). Als de betrouwbaarheid laag is omdat de groep studenten erg homogeen is, is dat geen reden de toets te diskwalificeren. Om die reden is er een alternatief geformuleerd voor het bepalen van het minimale aantal vragen dat voor een toets nodig is (voor de geïnteresseerde lezer: Wilcox, 1976; gaat uit van het binomiale model voor dichotome vragen met een aselecte trekking van items per student en/of geen verschil in moeilijkheid; binnen de item-respons theoretische benadering is een oplossing voorhanden; dit zou voor 'teacher made' toetsen echter te ver voeren).

Een andere toepassing vinden wij in de berekening van de (gemiddelde) variantie van de meetfouten en de wortel daaruit, de standaardmeetfout. De standaardmeetfout (zie formule II.9 in Bijlage II) wordt gebruikt in de cesuurmethode die door Wijnen (1971) werd voorgesteld en die besproken wordt in hoofdstuk 13.

Tabel 6. De gezamenlijke verdeling van twee parallelle toetsen

score op toets 2	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
30																			1		2	2	1	6	2
29																1	1	1		5	10	9	6	8	3
28											1				1	1	2	4	2	5	11	10	5	8	5
27																3	1	3	6	11	11	15	12	4	
26														1	1	4	4	7	11	9	7	6	9	6	
25								1					2	5	5	6	5	14	6	6	10	10	6	4	
24										1		2	1	4	5	1	8	8	7	9	6	6	5	3	1
23								1			2	7	1	3	4	4	13	14	4	6	6	3	3	2	
22								1			3	2	3	6	4	8	8	5	7	3	2		2	1	
21									1		4	7	5	7	7	9	8	8	4	3	1		2		
20					1	1	1	1	3		4	3	9	3	5	5	8	2	7	6	1			1	
19					1			1	1	1	3	8	7	6	8	7	5	6	2	1	3	2			
18								2	5	2	4	4	5	11	3	5	1	3	1	1	1				
17					1	2	4	1	6	2	3	1	2	4	7	2	4	1							
16			1	2	1	1	2	3	3	5	4	5	5	2	6	2	1		1						
15					1	1	5	3	4	6	6	5	2	1	3	2		1	1						
14	1				3		2	3	3	7	4	4	1	1	2										
13				1		2	2	4	2	1	4	2	4	5		1									
12			1	2	1		1	1	2	2	1	1													
11			3	2	1	3	1	1	3	5		1		1		1									
10			1	1	2	3		1	2		1														
9				2		2	1			2		1													
8	1		1		1	1								1											
7					1			1																	
6																									

score op toets 1

5. De cesuur zakken/slagen en de indeling van studenten in twee groepen

Scores worden omgezet in cijfers. Het onderscheid tussen een voldoende, bijvoorbeeld 6 of hoger, en een onvoldoende, 5 of lager, is het belangrijkst. Veronderstel dat wij bij de beide toetsen in Tabel 6 de cesuur voldoende/onvoldoende bij de score 19 leggen, dat wil zeggen de scores 19 en hoger zijn voldoende. Dan kunnen wij de gegevens in Tabel 6 samenklappen tot Tabel 7. Voor 166 studenten maakt het uit op grond van welke test een zak/slaagbeslissing zou worden genomen: 76 studenten slagen voor test 1, maar zakken voor test 2; 90 studenten slagen voor test 2, maar zakken voor test 1.

De consistentie van beslissingen hangt af van de betrouwbaarheid. Hoe hoger de betrouwbaarheid, des te hoger de consistentie van de beslissingen. De mate van consistentie hangt echter ook af van de dichtheid van de verdeling rond de cesuur. Als de cesuur dicht bij het toetsgemiddelde ligt, zal men in het algemeen meer inconsistente beslissingen mogen verwachten dan indien de cesuur ver onder het gemiddelde ligt.

Tabel 7. De consistentie van de beslissingen op twee parallelle tests

voldoende op test 2	90	618
onvoldoende op test 2	216	76
	onvoldoende op test 1	voldoende op test 1

Voor sommigen is een maat voor consistentie een alternatief voor een schatting van de betrouwbaarheid. De schatting van de consistentie van parallelle tests is moeilijker: wij hebben namelijk de beschikking over maar één toets. Ook deze maat zegt niet alles. Het percentage inconsistente beslissingen is hoog als de cesuur in een gebied valt met veel scores, zelfs bij een accurate toets. Veel van de inconsistente beslissingen betreffen dan een groep studenten met een ware score niet al te ver van de cesuur. Inconsistentie zegt niet alles over de nauwkeurigheid waarmee beslissingen zijn genomen. Dat hangt ook af van de adekwaatheid van de cesuur. Veronderstel dat wij de eisen veel te hoog hebben gesteld, met een te hoge cesuur. In dat geval zouden veel studenten *ten onrechte* zakken. Dat kunnen wij niet aan een tabel als Tabel 7 aflezen.

De cesuurbepaling zelf wordt in hoofdstuk 13 besproken.

6 Toetsen verschillen in moeilijkheidsgraad

Als wij twee verschillende toetsen afnemen, zijn de gemiddelde scores in het algemeen verschillend. Dat kan liggen aan het feit dat de toetsen aan twee verschillende groepen zijn voorgelegd, en die groepen kunnen qua niveau verschillen. Het is natuurlijk ook mogelijk dat de toetsen in moeilijkheid verschillen of dat zowel de toetsen als de groepen verschillen.

Het is gemakkelijk in te zien de toetsen in ieder geval niet volstrekt uitwisselbaar zijn qua moeilijkheid. Wij kunnen bijvoorbeeld een bestaande toets in twee helften delen, zonder dat wij eerst naar de *p*-waarden van de items kijken. Als wij dan het gemiddelde van beide toetsen (over dezelfde groep studenten) berekenen, dan zien wij dat de toetsen verschillen.

Als wij nu bij elke toets dezelfde cesuur gebruiken, dan geeft de schatting van de betrouwbaarheid een iets te mooi beeld van de nauwkeurigheid van de toetsing: er is immers geen rekening gehouden met de verschillen in moeilijkheidsgraad. Wij kunnen proberen na te gaan of een toets moeilijker of gemakkelijker is dan andere toetsen en de cesuur voor het geschatte verschil in moeilijkheidsgraad compenseren. Dat is het onderwerp van het hoofdstuk over normhandhaven (hoofdstuk 13).

7 Itemsteekproeven uit grote vragenbanken

Veronderstel, wij hebben een grote itemvoorraad opgebouwd en op een computer opgeslagen. Wij willen de studenten m.b.v. de computer tentamineren. Dat lijkt onder de omstandigheden heel efficiënt. Wellicht buiten wij de mogelijkheden van de computer uit door vragen op te nemen die wij nooit zo in een papieren versie hadden kunnen gebruiken.

Als wij de computer gebruiken, dan is het zinloos om een grote groep studenten op hetzelfde tijdstip op één locatie dezelfde toets te laten maken. Wij kunnen beter de studenten in kleinere groepen, eventueel individueel toetsen afnemen, vanaf één of verschillende locaties. Uiteraard kunnen wij niet dezelfde vragen stellen. De vragen – en de antwoorden – zouden snel uitlekken.

Wij moeten verschillende studenten verschillende toetsen aanbieden. Om de toetsing beheersbaar te houden laten wij de software ervoor zorgen dat elke student een andere toets krijgt. De vraagselectie kan op verschillende manieren plaatsvinden. Wij beperken ons hier tot twee verschillende methoden die gebaseerd zijn op aselecte trekking van vragen.

Bij de eerste methode wordt een aselecte steekproef van een bepaald aantal vragen getrokken uit een itembank: een grote voorraad vragen. Deze eenvoudige methode leidt tot het gebruik van het zogenaamde binomiale testmodel (zie Bijlage II); wij gaan nog steeds uit van items met een goed-fout score. In dit geval kunnen wij de betrouwbaarheid niet schatten met behulp van coëfficiënt alpha of KR20. KR21 (Formule II.7) geeft voor deze wijze van toetsing de schatting van de betrouwbaarheid. Er is een verschil in moeilijkheid van de items die de verschillende studenten te beantwoorden krijgen. Dat werkt door in de nauwkeurigheid waarmee wij de studenten onderling kunnen vergelijken. De berekende betrouwbaarheid zal iets lager uitvallen dan de betrouwbaarheid van een toets waarbij alle studenten dezelfde vragen moeten beantwoorden.

De methode van aselecte steekproeftrekking heeft nadelen. Als een bepaald onderwerp zich gemakkelijk voor het maken van vragen leent, is het mogelijk dat relatief veel items uit de itembank over dat onderwerp gaan. Dat onderwerp krijgt vervolgens teveel nadruk bij de toetsing. Als de items sterk in moeilijkheid verschillen, leidt de itemselectie bovendien tot toetsen die sterk in moeilijkheid verschillen, hetgeen gereflecteerd wordt in een lage betrouwbaarheid.

Wij brengen daarom een modificatie aan. De items worden volgens een toetsmatrijs gerubriceerd, ofwel in verschillende strata ingedeeld. Dat kan een indeling naar onderwerp zijn, maar tegelijk ook naar vraagniveau (bijvoorbeeld een indeling naar kennis versus meer-dan-kennis; voor een taxonomie van niveaus zie: Bloom, Engelhart, Furst, Hill & Krathwohl, 1956) en vraagtype (vierkeuze, waar-onwaar). Veronderstel nu dat elk item behoort bij één van de k verschillende strata. Nu stellen wij voor elke student een toets samen via de methode van de gestratificeerd aselecte steekproeftrekking. Dat houdt in: wij nemen aselect n_1 items uit de voorraad vragen voor stratum 1, n_2 items uit de voorraad vragen voor stratum 2, enz. Met een goede keuze van de aantallen te selecteren items n_1 tot en met n_k vermijden wij dat een onderwerp ten onrechte teveel of te weinig aandacht krijgt. Bij een goede indeling van de items is het bovendien waarschijnlijk dat

sommige strata relatief moeilijke items bevatten en andere relatief gemakkelijke. Daarmee wordt de variatie in moeilijkheid beperkt. De gestratificeerde methode levert een betrouwbaardere toetsing. Wij kunnen KR21 berekenen, in de wetenschap dat de coëfficiënt nu een onderschatting van de betrouwbaarheid geeft.[3]

Ook bij de (gestratificeerd) aselecte steekproefmethode van aanbieding kunnen wij itemindices berekenen al vergt dat een wat langere adem. Zo kunnen wij de *p*-waarde van een item berekenen als wij na de afname van een toets opslaan dat het desbetreffende item is afgenomen en wat het antwoord was. Wij kunnen de gegevens m.b.t. de totaalscores zo opslaan dat wij ook de item-totaalcorrelatie of de item-restcorrelatie kunnen berekenen. Wij moeten er wel rekening mee houden dat de r_{it} of r_{ir} bij de (gestratificeerd) aselecte steekproeftrekking iets lager uitvalt dan de r_{it} of r_{ir} bij een vaste toets voor een groep studenten. Dat komt omdat de totaalscores bij de (gestratificeerd) aselecte steekproeftrekking op steeds wisselende toetsen zijn gebaseerd.

[3] Een gestratificeerde variant van KR21 is mogelijk.

8. Open vragen: beoordelaarseffecten

Bij open vragen wordt meestal een beroep gedaan op een nakijker of beoordelaar. Dat brengt het risico met zich mee dat beoordelaarseffecten in het beoordelingsproces een rol gaan spelen. Meestal is de docent zich dat niet bewust, vooral niet als hij/zij de beoordelingen niet kan leggen naast die van een collega. Beoordelaarseffecten doen zich met name voor naarmate de vrijheid van beantwoording groter is (Coffman, 1971).

Er worden in de literatuur allerlei suggesties gedaan om beoordelaarseffecten terug te dringen. Zo kan men werken met modelantwoorden of met beoordelingsvoorschriften of een combinatie van beide. Bij uitgewerkte beoordelingsvoorschriften spreekt men van een analytische beoordeling. Tegenover de analytische beoordeling staat de globale beoordeling. Een beoordelaar kan het best een deel van het werk aan een voorlopig oordeel onderwerpen ten einde de regels voor het beoordelen scherper te kunnen vastleggen. Bij het beoordelen moet rekening gehouden worden met volgorde-effecten en vermoeidheid (zie verder blz 85-86 in Dousma, Horsten & Brants,1997).

Als er meer dan één beoordelaar beschikbaar is, kan men de beoordelaars op verschillende manieren inschakelen. De meest luxe situatie is die waarbij alle antwoorden op alle vragen door alle beschikbare beoordelaars worden nagekeken. Dit beoordelaarsschema wordt het gekruiste studenten × vragen × beoordelaars schema genoemd. Bij dit schema kunnen de verschillende beoordelaars hun beoordelingen onderling goed vergelijken. Wij kunnen nagaan in hoeverre de beoordelaars qua gemiddelde beoordeling (strengheid) en spreiding verschillen. Ook de correlatie tussen de beoordelingen van twee beoordelaars kan worden berekend.

Bij de schatting van de betrouwbaarheid kan met het aspect beoordelaars rekening worden gehouden. Als wij niet alleen over vragen willen generaliseren, maar ook over beoordelaars, is coëfficiënt alpha als schatting van de betrouwbaarheid ongeschikt. Voor de schatting van de betrouwbaarheid is een generalisatie van coëfficiënt alpha naar meer dan één aspect beschikbaar (De Gruijter & Van der Kamp, 2008).

Veelal is het standaard inschakelen van meer beoordelaars die onafhankelijk van elkaar een antwoord beoordelen, ondoenlijk. Het nakijkwerk wordt dan onder beoordelaars verdeeld. Er zijn in principe twee verschillende mogelijkheden: de studenten worden in groepen ingedeeld en elke beoordelaar kijkt het werkt van een groep na, of de vragen worden in groepen verdeeld en elke beoordelaar kijkt de antwoorden op een groep vragen na. In het eerste geval is informatie-uitwisseling bij een voorronde in de beoordeling mogelijk. Desondanks kunnen de beoordelaars qua strengheid blijven verschillen en kan het voor een student uitmaken wie zijn/haar antwoorden beoordeelt. In het tweede geval worden alle studenten die het tentamen hebben gedaan, op dezelfde wijze behandeld. de tweede methode verdient de voorkeur.

Nog een mogelijkheid is het gebruik van een nakijkersschema met overlap. Daarbij worden de studenten in groepen verdeeld en elke groep wordt door twee beoordelaars nagekeken. Elke beoordelaar beoordeelt twee groepen studenten. Op deze wijze kan worden nagegaan in hoeverre beoordelaars qua strengheid van elkaar verschillen. Het nakijkschema wordt schematisch in Schema 1 aangegeven.

	groep 1	groep 2	groep 3	groep 4
beoordelaar 1	XXX	XXX		
beoordelaar 2		XXX	XXX	
beoordelaar 3			XXX	XXX
beoordelaar 4	XXX			XXX

Schema 1. Een nakijkschema met overlap

9. Open vragen: verwerking

Open vragen nakijken is veelal handwerk. Veel docenten komen daardoor niet meer aan het invoeren van gegevens over de afzonderlijke vragen in een bestand toe. Dat is een gemiste kans (De Gruijter, 2000). Het is vrij eenvoudig om beoordelingen in een spreadsheet in te voeren. Het voordeel is dat het berekenen van totaalscores en het omzetten van de totaalscores in cijfers vrij eenvoudig en foutloos kan gebeuren. Bovendien kan heel snel een item- en toetsanalyse worden gedaan. Dat wordt toegelicht aan de hand van het voorbeeld in Schema 2.

	A	B	C	D	E	F	G	H	I	J	K	L
1	nr student	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	totaal
2	1	2	2	0	2	2	1	1	2	1	0	13
3	2	2	2	0	2	0	0	2	2	0	2	12
4	3	2	2	0	2	2	0	2	2	0	0	12
5	4	2	2	0	2	2	0	2	2	0	0	12
6	5	0	2	1	0	0	2	1	2	1	0	9
7	6	2	2	0	2	1	1	2	2	0	1	13
8	7	2	2	0	2	0	0	2	2	1	2	13
9	8	1	0	0	1	0	1	1	0	1	0	5
10	9	2	2	0	2	2	1	2	2	0	0	13
11	10	2	2	2	2	2	1	2	2	1	2	18
12												
13	M=	1.70	1.80	0.30	1.70	1.10	0.70	1.70	1.80	0.50	0.70	12.00
14	var=	0.46	0.40	0.46	0.46	0.99	0.46	0.23	0.40	0.28	0.90	10.89
15	s=	0.67	0.63	0.67	0.67	0.99	0.67	0.48	0.63	0.53	0.95	3.30
16	Mmax	2	2	2	2	2	2	2	2	2	2	
17	M/Mmax=	0.85	0.90	0.15	0.85	0.55	0.35	0.85	0.90	0.25	0.35	
18	rit=	0.65	0.75	0.45	0.65	0.58	-0.20	0.63	0.75	-0.13	0.53	
19	rir=	0.50	0.64	0.26	0.50	0.32	-0.38	0.53	0.64	-0.28	0.28	
20											alpha=	0.599

Schema 2. Spreadsheet met gegevens van een hypothetisch tentamen met open vragen

In de rij achter 'M' staat de gemiddelde score op de vragen en de gemiddelde totaalscore. Achter 'var' staan de varianties In de rij daaronder staat de standaardafwijking *s*. Achter 'M/Mmax' staat de gemiddelde score op een vraag (*M*) gedeeld door de maximale score (Mmax). Deze index geeft een indicatie van de gemakkelijkheid van de vraag en is vergelijkbaar met de *p*-waarde bij meerkeuzevragen met goed/fout scoring. In het voorbeeld hebben alle vragen een maximale score gelijk aan 2.

De r_{it} is, evenals *M*, de variantie en de standaardafwijking, berekend via een spreadsheetfunctie. Voor de berekening van r_{ir} moet meer moeite worden gedaan. Men kan het beste de volgende formule invoeren:

$$r_{ir} = \frac{s_t r_{it} - s_i}{\sqrt{s_t^2 - 2 s_i s_t r_{it} + s_i^2}},$$

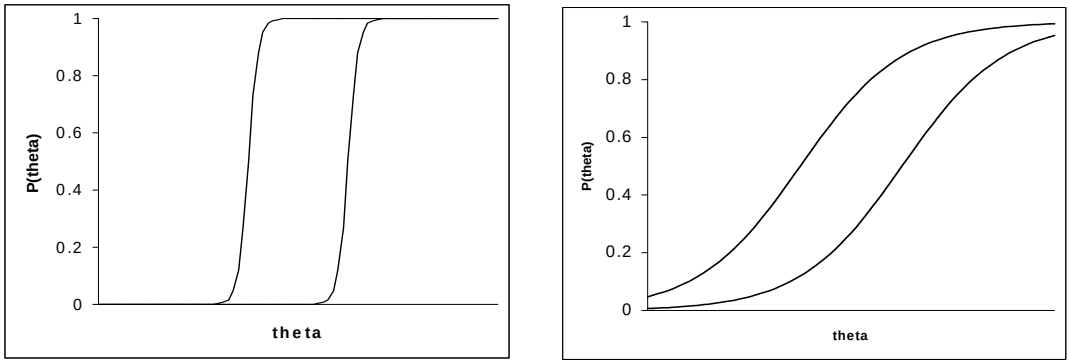
waarbij s_t de standaardafwijking van de totale toets is, en s_i de standaardafwijking van item *i*. Voor coëfficiënt alpha dient men eveneens een formule in te voeren (zie Bijlage II, formule II6). Een opzet zoals in Schema 1 wordt gegeven, is zo nodig via een macro aan te maken.

Indien de vragen worden gewogen, moeten de zaken iets anders worden aangepakt. daartoe staan twee wegen open. De eerste mogelijkheid is om het scoringsvoorschrift aan te passen. Wij kunnen bijvoorbeeld afspreken dat wij bij een vraag met gewicht 2 in plaats van een score 1 een score 2 geven. De tweede mogelijkheid is aanpassing van de formules voor het berekenen van de totaalscores, de item-restcorrelatie en coëfficiënt alpha.

Ten slotte, bij veel tentamens moet een open-vragengedeelte worden gecombineerd met een gesloten toets. Voor de berekening van de betrouwbaarheid van het totale tentamen kan (zie Bijlage II, formule II.8) worden gebruikt.

10 Wat is de optimale moeilijkheidsgraad van een vraag?

Als wij zo goed mogelijk onderscheid willen maken tussen de studenten lijkt het nuttig om vragen van een verschillende moeilijkheid in de toets op te nemen. De moeilijke vragen zijn dan vooral bedoeld om onderscheid te maken tussen de goede studenten, de gemakkelijke vragen discrimineren vooral tussen de minder goede studenten. Met vragen die sterk discrimineren is dat zo. Dat is in Figuur 2.a te zien. In de figuur staat θ voor de vaardigheid van de student en $P(\theta)$ geeft de kans aan dat een student met de desbetreffende vaardigheid het item goed beantwoordt. Deze kans loopt in de figuur sterk op met de vaardigheid. In de praktijk is het discriminerend vermogen van de afzonderlijke vragen dikwijls minder sterk. Figuur 2.b geeft daarvan een beeld. Dan is het beter om vragen te hebben die niet zoveel in moeilijkheid van elkaar verschillen.



(a) sterk discriminerende items (b) redelijk discriminerende items

Figuur 2. Discriminatie van items

Bij toetsing in het onderwijs gaat het er echter praktisch nooit om dat er binnen een groep studenten zo goed mogelijk moet worden gediscrimineerd. Het belangrijkste aspect van de summatieve toetsing is een zo accuraat mogelijke beslissing over het al dan niet laten slagen van studenten. In dat geval moeten de items zo nauwkeurig mogelijk zijn bij een vaardigheid rond het omslagpunt tussen een voldoende en een onvoldoende vakbeheersing. Als gokken geen rol speelt, moet de kans op een goed antwoord op dat niveau ongeveer een half zijn. Met gokken moet de kans ongeveer op het niveau

$$c + \tfrac{1}{2}(1-c)$$

liggen, waarbij c de gokkans is. Voor het gemak veronderstellen wij dat de gokkans gelijk is aan 1 gedeeld door het aantal antwoordalternatieven (de gokkans die wij bij empirisch onderzoek vinden, kan daarvan afwijken). Bij vierkeuzevragen zou de kans ongeveer 0.63 moeten zijn. Als wij ervan uitgaan dat meer dan de helft van de groep studenten die de toets maakt, geschikt is, moet de p-waarde van een vraag met een optimale moeilijkheid in de groep in ieder geval hoger dan 0.63 zijn.

11. Het optimale aantal alternatieven en gokken

Veel meerkeuzetoetsen bestaan uit vierkeuze-items. Dit vraagtype is in Nederland vooral door het werk van De Groot en Van Naerssen (1969) populair geworden. Waar-onwaarvragen kunnen echter ook heel goed zijn. Een nadeel bij waar-onwaarvragen is dat een ware stelling absoluut juist moet zijn. Studenten kunnen problemen met het beantwoorden van waar-onwaarvragen hebben, met als gevolg dat antwoordtendenties een rol gaan spelen (Grosse en Wright, 1985). Ebel (1982) stelt daarom voor met contrasten te toetsen. Hij geeft een voorbeeld. De vraag

An eclipse of the sun can only occur when the moon is new (T/F)

of

An eclipse of the sun can only occur when the moon is full (T/F)

kan vervangen worden door

An eclipse of the sun can only occur when the moon is 1) new 2) full.

Tweekeuze-vragen lijken in het algemeen een tweede nadeel te hebben tegenover vragen met meer dan twee alternatieven. Bij tweekeuze-vragen is de gokkans namelijk erg hoog, een half, en men zou dus mogen verwachten dat een toets met tweekeuze-vragen onbetrouwbaarder is.

Vanwege de grote gokkans wordt vooral bij tweekeuze (waar-onwaar) vragen wel eens een gokcorrectie toegepast. Bij een vraag met k alternatieven is de scoring met gokcorrectie:

1 punt	bij een goed antwoord
0 punten	als de student de vraag heeft overgeslagen
$-1/(k-1)$ punten	bij een fout antwoord, waarbij k het aantal vraag-alternatieven is.

Bij tweekeuze-vragen zijn de mogelijke scores dus 1, 0 en –1. Zo wordt gestimuleerd dat een student die het antwoord niet weet, de vraag niet beantwoordt. Dit moet leiden tot betrouwbaarder metingen. Daarmee zou het nadeel van een grote gokkans bij vragen met een klein aantal alternatieven voor een deel kunnen worden opgeheven.

De genoemde gokcorrectie is in het algemeen niet te verdedigen. Een student die het antwoord niet weet, kan de vraag overslaan en krijgt daarmee 0 punten. Een student die het antwoord gokt, heeft bij een tweekeuze-vraag 50 procent kans op het goede antwoord. Zijn verwachte score op de vraag is 0. Gokken wordt met de gokcorrectie dus niet echt tegengegaan. In het algemeen is het beter om goed-fout scoring te gebruiken en de studenten te instrueren om zoveel mogelijk een antwoord te geven dan om met een gokcorrectie te werken. Een uitzonderingssituatie is die waarbij studenten met weinig kennis in tijdnood dreigen te komen bij het beantwoorden van vragen op gebieden waarover zij enige kennis hebben. Dat is onder andere het geval bij de zogenaamde voortgangstoets die enkele opleidingen kennen. Bij de voortgangstoets krijgen de studenten van alle cursusjaren dezelfde lange toets met vragen die de totale leerstof bestrijken.

Als een scoringsformule van weinig nut is, lijkt het voor de hand te liggen om gokken zoveel mogelijk tegen te gaan door meer antwoordalternatieven bij een vraag op te nemen. Ook dat is niet zonder meer waar. Laten wij een driekeuzevraag nemen en daar een vierkeuzevraag van maken door het toevoegen van een alternatief. Als dat alternatief met veel moeite gevonden wordt, bestaat het risico dat het alternatief niet als afleider werkt. Dan blijft de vraag qua gokkans een driekeuzevraag.

Als wij wel een goede afleider vinden, daalt de gokkans daadwerkelijk. De vierkeuzevraag kan echter wel tot een langere leestijd voor de studenten leiden. In een tentamen met een vaste tijdsduur kunnen wij meer driekeuzevragen opnemen dan vierkeuzevragen. Het grotere aantal driekeuzevragen compenseert de grotere onnauwkeurigheid vanwege het geringere aantal alternatieven. Dat geldt vooral bij de bepaling van de kennis en vaardigheden van niet al te slechte studenten. Toetsen met driekeuzevragen en tweekeuzevragen kunnen daarom beter zijn dan toetsen met vierkeuzevragen. En als een deel van de stof zich beter leent voor bevraging met waar-onwaar vragen en een ander deel voor bevraging door bijvoorbeeld vierkeuzevragen, ligt het voor de hand om beide vraagvormen in de toets te gebruiken.

12. Weging van vragen en vraagalternatieven

Wij gaan uit van de volgende meerkeuzevraag:

Stelling I: toetsen met meerkeuzevragen zijn *niet* geschikt om inzicht te toetsen
Stelling II: de score van een fout antwoord bij de correctie voor gokken bedraagt – 0.5 bij een driekeuzevraag.

a. I is waar, II is waar
b. I is waar, II is onwaar
c. I is onwaar, II is waar
d. I is onwaar, II is onwaar.

Alternatief c is het juiste alternatief; de alternatieven a en d zijn niet geheel onjuist. Wij zouden kunnen afspreken bij antwoord c 1 punt te geven, bij antwoord b 0 punten en bij a of d een half punt. Dan wegen wij de verschillende alternatieven. Wij zouden ook kunnen afspreken antwoord c twee punten te geven en a of d 1 punt. Dan wegen wij ook de vraag: het verschil tussen de maximale en de minimale score op de vraag is 2 punten i.p.v. 1 punt. In het voorbeeld zijn twee tweekeuzevragen samengevoegd om een vierkeuzevraag te krijgen. Het is beter – minder gekunsteld – om de twee waar-onwaarvragen als zodanig op te nemen:

Vraag 1. Toetsen met meerkeuzevragen zijn *niet* geschikt om inzicht te toetsen
a. waar
b. onwaar

Vraag 2. De score van een fout antwoord bij de correctie voor gokken bedraagt – 0.5 bij een driekeuzevraag.
a. waar
b. onwaar

De scoring van de vierkeuzevraag waarbij aan het goede antwoord 1 punt wordt toegekend, komt overeen met een weging van de waar-onwaar vragen met een half. Bij beide waar-onwaarvragen wordt in dat geval een half punt voor het goede antwoord gegeven. Een dergelijk klein gewicht voor een waar-onwaar vraag in een toets waarin ook vierkeuzevragen met de score 0 voor een fout antwoord en 1 voor een goed antwoord voorkomen, is op psychometrische gronden niet te verdedigen. Als de waar-onwaar vragen niet te moeilijk zijn is de scoring 1 voor een goed antwoord en 0 voor een fout antwoord ook bij tweekeuzevragen te verdedigen.

In het algemeen kent men twee soorten wegingen. De eerste soort is een weging op a priori overwegingen. Men kan een vraag zwaarder laten wegen om het desbetreffende onderwerp een groter gewicht in de totaalscore toe te kennen. In het algemeen is het echter beter om over een onderwerp dat van belang is een redelijk aantal vragen op te nemen. Bij open vragen kan een a priori weging wel nuttig zijn. Sommige vragen vergen meer tijd voor de beantwoording en dat kan men door een groter gewicht laten meetellen. De maximaal haalbare score op de vragen moet tevoren aan de studenten worden meegedeeld zodat zij daar bij de beantwoording rekening mee kunnen houden.

De tweede manier van wegen is psychometrisch van aard. Men zoekt daarbij gewichten die de nauwkeurigheid van de gewogen totaalscore verhoogt. De nauwkeurigheid van de berekening van de optimale gewichten is afhankelijk van de omvang van de groep studenten waarvan de gegevens worden geanalyseerd. De gewichten worden in ieder geval de eerste keer achteraf pas vastgesteld en zeggen de studenten weinig of niets.

Indien de afzonderlijke vragen verschillend worden gewogen met gewichten w_i, dan moeten wij de betrouwbaarheidsschatting aanpassen. Veronderstel dat wij beschikken over de variantie van de gewogen totaalscores en de ongewogen varianties van de afzonderlijke vragen. Dan moeten wij bij de berekening van coëfficiënt alpha in de noemer de variantie van de gewogen totaalscores gebruiken. In de teller vervangen wij de som van de item varianties door een gewogen som van de item varianties met gewichten w_i^2.

13. Cesuurmethoden: normstellen en normhandhaven

13.1. Inleiding

De beslissing omtrent de cesuur onvoldoende/voldoende is een heel belangrijke. Het is bovendien een moeilijke beslissing. Soms wordt gekozen voor schijnbare eenvoud. Een docent gaat er bijvoorbeeld van uit dat een student de helft van de toetsvragen goed moet kunnen beantwoorden. Bij een meerkeuzetoets zou de cesuur bij 50 procent van het aantal vragen kunnen liggen, ware het niet dat de studenten ook door gokken vragen goed kunnen beantwoorden. Een gokcorrectie is dus op zijn plaats.

Laten wij het probleem wat algemener benaderen. Wij schrijven p_c voor de proportie items die de student goed moet kunnen beantwoorden. Als alle vragen k alternatieven hebben en n het aantal toetsvragen is, dan kan de cesuur C geschreven worden als:

$$C = n[p_c + c(1 - p_c)],$$

waarbij $c = 1/k$ de theoretische gokkans is.

De zo vastgestelde cesuur is een *absolute* cesuur. De zak/slaaggrens is vastgesteld zonder kennis te nemen van de prestaties van de studenten. Voordeel is dat de norm tevoren aan de studenten bekend kan worden gemaakt (met een slag om de arm). Het voordeel is extra groot als er een vergelijkbaar proeftentamen of voorbeeldtentamen beschikbaar is. Dan hebben de studenten concrete informatie over het niveau dat van hen wordt verwacht.

Er is ook een – groot – nadeel aan de methode. De methode levert een arbitraire cesuur op tenzij de moeilijkheid van de items adequaat is afgestemd op het niveau op de grens tussen *voldoende* en *onvoldoende*. Bij een onverwacht hoog percentage gezakten zou de docent dan ook kunnen concluderen dat de items toch te moeilijk waren en bijgevolg de cesuur willen aanpassen. Soms is die aanpassing eenvoudig: als een afleider bij nader inzien niet zo een slechte keus is, kan dit alternatief ook goed worden gerekend en daarmee kan het percentage geslaagden iets omhoog. Het risico is echter groot dat voor de aanpassing achteraf geen goede argumenten kunnen worden gegeven.

13.2. Normstellen en normhandhaven

De onzekerheid over de correctheid van de cesuur voldoende/onvoldoende is ongetwijfeld het grootst bij de eerste tentamengelegenheid die wordt georganiseerd. Als het tentamen slechter gemaakt is dan verwacht, kan de docent op goede gronden besluiten de cesuur te laten zakken. Wij komen daar nog op terug. De opgedane ervaring is relevant voor toekomstige tentamens. De docent kan besluiten ook in de toekomst een lagere cesuur te gebruiken of gemakkelijker items te construeren. Veel docenten zullen dan minder geneigd zijn om de cesuur aan te passen aan de resultaten van de studenten. De resultaten kunnen namelijk niet alleen meevallen of tegenvallen omdat het tentamen te gemakkelijk of te moeilijk is, maar ook omdat de groep studenten relatief beter of slechter is. Het is daarom goed om bij de keuze tussen cesuurmethoden een onderscheid te maken

tussen de bepaling van de cesuur bij de eerste tentamengelegenheid, de normstelling, en de bepaling bij latere gelegenheden waarbij veelal naar het handhaven van eenmaal vastgestelde normen wordt gestreefd. Er is één methode, voorgesteld door De Gruijter, die alleen bedoeld is voor de fase van normstelling (De Gruijter, 1985). Voor normhandhaving moet een beroep op een andere methode worden gedaan. Andere methoden, zoals die van Nedelsky (1954) en Angoff (1971) kunnen zowel bij de eerste gelegenheid als bij latere gelegenheden worden toegepast, maar dat hoeft men niet te doen. Bij weer andere methoden, zoals de door Wijnen (1971) voorgestelde methode, is normhandhaving niet het doel van de cesuurbepaling.

13.3. Normhandhaving, een vaste cesuur en een vast percentage gezakten

Als wij voor elk tentamen veel vragen met een goede steekproefmethode uit een grote voorraad vragen selecteren, verkrijgen wij tentamens die elkaar in moeilijkheid niet veel ontlopen. Dan kan men het beste steeds dezelfde cesuur gebruiken. Hier voldoet de meest eenvoudige absolute methode om de norm te handhaven.

Als het samenstellen van de toets minder systematisch gebeurt, heeft de samensteller van de toets geen controle over de moeilijkheid. Het is mogelijk dat de ene toets veel moeilijker is dan de andere. Dan zou een vaste cesuur als consequentie hebben dat de hoogte van de eisen voor verschillende groepen studenten feitelijk ongelijk zijn.

Bij de meest eenvoudige relatieve methode houden wij het percentage gezakte studenten constant. Indien de groepen studenten groot zijn en van jaar tot jaar vergelijkbaar, dan behoort bij elk tentamen (afgezien van herkansingstentamens) een ongeveer even groot percentage studenten te zakken. Normhandhaving is in dit geval te verwezenlijken door die cesuur te nemen die hetzelfde percentage gezakte studenten oplevert: een relatieve cesuurmethode.

Als de groepen studenten en de tentamens redelijk vergelijkbaar zijn, maar de aantallen studenten en vragen niet zo groot dat vaste cesuur of een vast percentage gezakten een adequate oplossing voor het cesuurprobleem is, kan men een tussenweg tussen een vaste cesuur en een vast percentage gezakten trachten te vinden. Men zou kunnen denken aan een correctie van de scores op een tentamen afhankelijk van de prestaties van de groep studenten en de omvang van de groep studenten.

13.4. Cesuurmethoden

Er zijn veel methoden voor het bepalen van de cesuur voorgesteld, een uitvloeisel van het probleem dat het moeilijk is om een voor alle betrokkenen acceptabele oplossing te bedenken. Wij zullen maar een aantal methoden beschrijven. Voor een overzicht van methoden, zie Hambleton & Pitoniak (2006).

De te bespreken cesuurmethoden kunnen wij in drie groepen uiteen laten vallen: absolute methoden,relatieve methoden en compromismethoden. De methoden zijn:

1. absolute methoden
   a vaste cesuur
   b methode van Nedelsky
   c methode van Angoff
   d equivalering
2. relatieve methoden
   a methode Wijnen
3 compromismethoden
   a methode van Hofstee
   b methode van De Gruijter
   c methode van Cohen-Schotanus, Van der Vleuten & Bender

De vaste cesuur

Stel a priori op basis van de aard van de vragen een cesuur vast, zoals 50 procent goed plus een gokcorrectie bij meerkeuze vragen en waar/onwaar vragen. De methode is nogal arbitrair. Men kan overwegen om de eerste cesuur te berekenen door toepassing van een andere absolute methode of een compromismethode en de verkregen cesuur te blijven toepassen bij volgende tentamens.

De methode van Nedelsky (1954)

De methode van Nedelsky is geschikt voor meerkeuzevragen. Nedelsky stelde voor een hypothetische student met een prestatieniveau op de grens tussen voldoende en onvoldoende voor ogen te nemen. Bij elke meerkeuzevraag kan de docent beoordelen welke alternatieven deze student als fout moet kunnen afstrepen. Verondersteld wordt dat de student onder de resterende alternatieven gokt. Wij nemen een voorbeeld. Stel dat de student bij een vierkeuzevraag twee alternatieven kan wegkruisen. Tussen de twee overgebleven alternatieven wordt gegokt. De kans op de keuze voor het goede alternatief, en dus de verwachte score op het item, bedraagt een half (één gedeeld door het aantal alternatieven). De kansen worden voor alle items berekend en vervolgens opgeteld. De uitkomst geeft de cesuur.

De methode van Nedelsky is een subjectieve methode. De uitkomst is afhankelijk van de beoordelaar. Verschillende beoordelaars kunnen tot verschillende uitkomsten komen (als er meer dan één beoordelaar is, kunnen wij de door hen berekende cesuren natuurlijk middelen). De methode is ook absoluut: de cesuur wordt vastgesteld zonder rekening te houden met de resultaten van de studenten. De methode kan als normstellingsmethode worden gebruikt of onderdeel vormen van de normstelling met een compromissmethode. De methode kan ook na de eerste tentamengelegenheid worden toegepast. Het is echter onder bepaalde condities mogelijk dat voor normhandhaving de vaste cesuur te prefereren is.

De methode van Angoff (1971)

Angoff heeft een aantal voorstellen voor de cesuurbepaling bij meerkeuzetoetsen gedaan. De meest bekende vorm is die waarbij de beoordelaar bij elke vraag schat hoeveel procent van een groep grensstudenten (studenten met een prestatieniveau op de grens voldoende/ onvoldoende) de vraag goed beantwoordt. De cesuur is de som van deze kansen. Bij deze methode kunnen dezelfde opmerkingen gemaakt worden als bij de methode Nedelsky.

Equivalering

Equivalering is het vertalen van de schaal van de ene toets naar die van een andere. Equivalering kan dus gebruikt worden om eenmaal vastgestelde normen te handhaven. Equivalering is onderwerp van het volgende hoofdstuk.

De methode van Wijnen (1971)

De methode van Wijnen is een relatieve methode. De resultaten van de groep studenten die tentamen heeft gedaan, bepalen de hoogte van de cesuur. De gedachte achter de methode is dat het groepsgemiddelde representatief is voor het niveau dat de studenten gegeven de omstandigheden konden bereiken. Wie teveel aan de onderkant van dat gemiddelde afwijkt, is gezakt. De methode is niet bedoeld voor tentamens met herkansers.

Voor de cesuurbepaling volgens Wijnen is de standaardmeetfout (zie Bijlage II) nodig, en dus een schatting van de betrouwbaarheid $r_{xx'}$. De cesuur is gelijk aan het gemiddelde min tweemaal de standaardmeetfout, in formulevorm:

$$C = \overline{X} - 2s_X\sqrt{1-r_{XX'}}\ .$$

De methode van Hofstee (1977)

Als bij een absolute cesuur een hoger percentage studenten zakt dan je redelijkerwijze had mogen verwachten ligt het voor de hand de cesuur te verlagen. Wat je dan doet, is achteraf relatieve informatie bij het bepalen van de cesuur gebruiken. Hofstee geeft aan dat je beter tevoren duidelijkheid kan scheppen over de mate van aanpassingen waartoe je, gegeven de prestaties van de studenten bereid bent. Hij stelt een compromismodel voor.

Voor de beschrijving van de methode van Hofstee wijken wij enigszins af van de beschrijving in het oorspronkelijke artikel. De methode werkt als volgt:

- Stel de cesuur vast die je zou willen toepassen: c_0 (het is het handigst om de cesuur uit te drukken als percentage van de maximaal haalbare score[4]).
- Stel het percentage gezakten vast waarbij je deze cesuur adequaat vindt: f_0.
- Geef aan in welke mate je de cesuur wilt aanpassen als het percentage gezakten van f_0 afwijkt. Dit resulteert in de formule:

  $c = c_0 + a(f - f_0)$ (ook te schrijven als $f = f_0 + (c - c_0)/a$)

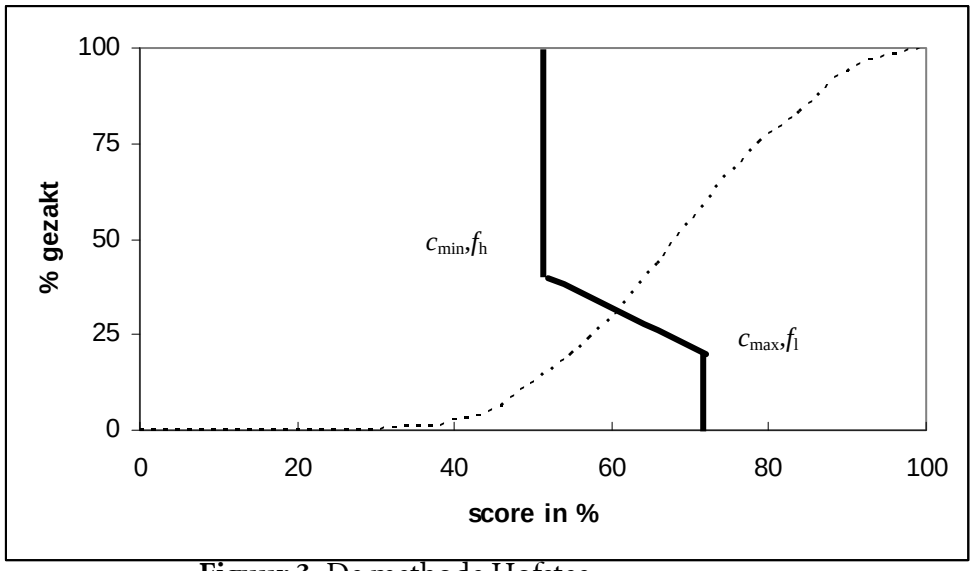
  waarbij a een negatief getal is.
- Stel de laagste cesuur vast die acceptabel is: c_{min}.
- Stel de hoogste cesuur vast: c_{max}.
- Bij het tentamen kunnen wij voor elke combinatie van cesuur en het daarbij resulterend percentage gezakten nagaan of aan de formule is voldaan. In de praktijk voldoet geen enkele combinatie exact aan de

[4] Hofstee werkt in verband met de mogelijkheid tot gokken bij meerkeuzetoetsen niet met een scorepercentage, maar met een kennispercentage.

formule. Wij zoeken dus een combinatie van cesuur en het daarbij behorend percentage gezakten waarvoor de absolute waarde van $c - a(f - f_0) - c_0$ zo klein mogelijk is.

- Als de cesuur daarvoor onder c_{min} zou uitkomen, houden wij c_{min} aan als absolute ondergrens voor de cesuur.
- Als de cesuur boven de c_{max} uitkomt, houden wij c_{max} aan als cesuur (de methode is ook te schrijven als een methode waarbij een cesuur c_{max} alleen naar beneden wordt bijgesteld).

De methode kan in een Figuur worden neergezet. Uit de figuur is te zien dat de procedure ook vastligt als de punten (c_{max},f_l) en (c_{min},f_h) zijn bepaald. In Figuur 3 geeft de dikke lijn de beslisregel aan, met als uiterste grenzen voor c c= 52 en c =72, bij een percentage gezakte studenten van 40, resp. 20. De gestippelde lijn geeft voor een bepaald tentamen aan hoeveel procent van de studenten zou zakken als een bepaalde cesuur gekozen zou worden. Bij een cesuur van 60 zou 32 procent van de studenten zakken. Dit is tevens de combinatie van cesuur en percentage gezakte studenten die de methode Hofstee in dit voorbeeld aangeeft: bij deze combinatie is het verschil tussen de dikke lijn en de stippellijn het kleinst, m.a.w. bij deze combinatie ligt het empirisch resultaat het dichtst bij wat de compromismethode aangeeft.
In de communicatie met de studenten kan de dikke lijn uit de figuur uitgangspunt zijn. Als benadrukt wordt wat de cesuur maximaal is, kan elke lagere cesuur als een 'meevaller' worden beschouwd.



Figuur 3. De methode Hofstee

De methode van Hofstee is heel goed als normstellingsmethode toe te passen. De methode kan niet worden toegepast op herkansingen.

De methode van De Gruijter (1985)

De methode van De Gruijter gaat eveneens uit van een combinatie van c (als percentage van de maximale score) en f die optimaal is: c_0, f_0. De docent is niet volstrekt zeker over de waarde van c_0. De onzekerheid wordt uitgedrukt in termen van een verdeling rond c_0 met een standaardafwijking gelijk aan s_c. De docent is

eveneens onzeker over de waarde van f_0. Dat wordt uitgedrukt in termen van een verdeling rond f_0 met een standaardafwijking gelijk aan s_f. Alleen de verhouding tussen de twee standaardafwijkingen $r = s_f / s_c$ is nodig.

Wij zoeken de combinatie van cesuur en percentage gezakten waarvoor de absolute waarde van

$$r^2(c - c_0)^2 + (f - f_0)^2$$

minimaal is.

De methode is door De Gruijter voorgesteld als normstellingsmethode, een methode die eenmalig wordt toegepast.

De methode van Cohen-Schotanus, Van der Vleuten & Bender (1996).

Deze methode gaat uit van een in principe absolute normering: de studenten moeten *k* procent van het maximaal aantal punten halen voor een voldoende. Om daarvoor te corrigeren wordt een relatief referentiepunt in de normbepaling meegenomen. Wij nemen daarbij niet *k* procent van de maximale score, maar *k* procent van bijvoorbeeld de score die door de student op de 95ste percentiel wordt behaald.

Bij meerkeuze-items zou *k* als kennispercentage kunnen worden vastgesteld. dan is een gokcorrectie nodig. Als het tentamen n vragen bevat, de student op de 95ste percentiel n* vragen goed heeft, c de gokkans is, en p de vereiste proportie kennis (k/100), dan wordt de gecorrigeerde cesuur:

Cesuur = $nc + (n^* - nc)p$.

In de praktijk krijgt men altijd een verlaagde cesuur. De methode grijpt niet rechtstreeks op het percentage gezakte studenten in. De methode is gevoelig voor steekproeffluctuaties. Evenals de andere relatieve methoden en compromismethoden is deze wijze van cesuurbepaling niet geschikt voor herkansingen.

14 Van score naar cijfer

Als de grens voldoende/onvoldoende bekend is kunnen cijfers gegeven worden. Als wij 10 hele cijfers (van 1 t/m 10) gebruiken, moeten wij de range van mogelijke scores in 10 intervallen verdelen. Bij meerkeuzetoetsen willen wij alle scores onder de gokkans in ieder geval het laagste cijfer geven. Verder zou het gemakkelijk zijn als wij de scorerange voor de cijfers in gelijke intervallen zouden kunnen verdelen. Dat lukt natuurlijk echter niet altijd vanwege de positie van de cesuur. Wat wij wel kunnen doen is de scores boven de cesuur in min of meer gelijke intervallen te verdelen ten behoeve van de becijfering. De scores onder de cesuur worden apart in cijferintervallen ingedeeld. Wij geven een voorbeeld met hele cijfers waarbij door afronding ook scores onder de maximale score een 10 kunnen krijgen.

Notatie:
C = cesuur
Max = maximale score
g = gokscore
S = score

Als $S < C$ dan cijfer = maximum van 1 en
AFRONDEN tot heel getal $5.499 - 4.5 \times (S - C)/(g - C)$
Als $S \geq C$ dan cijfer is
AFRONDEN tot heel getal $5.5001 + 4.5 \times (S - C)/(Max - C)$

De getallen 5.499 en 5.5001 zijn hier in de formules opgenomen in plaats van de waarde 5.5 om eventuele afrondingsproblemen bij exact de waarde 5.5 te voorkomen.

15 Equivaleren en item-respons theorie

Als wij meer dan één toets hebben afgenomen, worden wij geconfronteerd met de mogelijkheid dat de toetsen qua moeilijkheid van elkaar kunnen verschillen. Een indicatie daarvan kan zijn dat de ene toets slechter is gemaakt dan de andere. Het probleem is dat de slechtere prestaties wellicht niet veroorzaakt zijn door een moeilijker toets, maar door een slechtere groep studenten. Veronderstel nu dat één vraag in beide toetsen is opgenomen. Deze vraag is bij beide gelegenheden even relevant. Laten wij er bovendien vanuit gaan dat de vraag na de eerste gelegenheid niet bekend is geraakt onder de studenten. Als nu deze vraag even goed gemaakt wordt in de goed gemaakte toets als in de slecht gemaakte toets, dan lijkt het erop dat het verschil in prestaties op beide toetsen niet veroorzaakt is door een verschil in niveau van de twee groepen studenten.

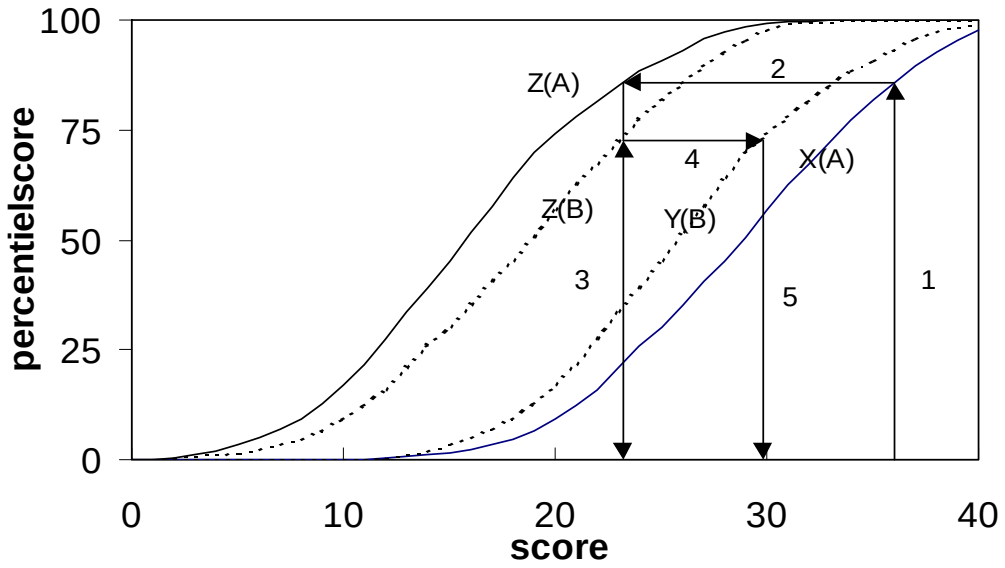
Bij het equivaleren van toetsen wordt een aantal gemeenschappelijke items gebruikt om eventuele verschillen tussen toetsen te bepalen en scores te berekenen voor de ene toets die equivalent zijn met bepaalde scores op de andere toets. Zo kan voor een nieuwe toets de score berekend worden die equivalent is met de zak/slaaggrens op een oude toets. Verschillende methoden voor equivaleren zijn voorgesteld. Helaas wordt in de literatuur niet altijd duidelijk dat toepassing van equivaleringsmethoden de uiterste voorzichtigheid vergt. En nog in 1997 worden in het boek van Dousma, Horsten en Brants 'equivaleringsmethoden' (de lineaire of niet-lineaire normhandhavingsmethode) uit de zestiger jaren behandeld die psychometrisch niet te verdedigen zijn.[5]

Wij introduceren het onderwerp *equivaleren* met een voorbeeld waarbij er geen gemeenschappelijke items zijn, maar een apart afgenomen toets die door beide groepen studenten is gemaakt. Veronderstel dat wij twee toetsen *X* en *Y* hebben die wij willen vergelijken. Toets *X* is afgenomen bij een grote groep studenten, laten wij zeggen groep A. Toets *Y* is later afgenomen bij een tweede grote groep studenten, groep B. Beide groepen studenten maken ook een derde toets, toets *Z*. Toets *Z* meet dezelfde kennis en vaardigheden als de toetsen *X* en *Y*, en is even betrouwbaar als de twee andere tests. Dankzij de gemeenschappelijke toets *Z* kunnen wij voor toets *Y* scores berekenen die equivalent zijn aan bepaalde scores van toets *X* (zie Figuur 4). Eén van de mogelijkheden is de volgende. Bij toets *X* haalt 85 procent van de studenten uit groep A een score 36 of minder. Binnen dezelfde groep haalt 85 procent van de studenten een score van 23 of minder bij toets *Z*. Een score van 36 op *X* komt dus overeen met een score van 23 op toets *Z*. Binnen groep B kunnen wij nagaan welke scores op *Y* overeenkomen met scores op *Z*. Tweeënzeventig procent van de studenten in groep B heeft een score van 23 of minder op toets *Z* en eveneens 72 procent heeft een score van 30 of lager op *Y*. Dan correspondeert een score van 23 op *Z* met een score van 30 op *Y*. Dus een score van 36 op *X* correspondeert met een score van 30 op *Y*.

Via equivaleren kunnen wij voor alle scores op toets *Y* aangeven met welke score op *X* zij equivalent zijn. Zo komen wij ook te weten welke score op toets *Y* overeenkomt met de grens voldoende/onvoldoende op toets *X*. Wij kunnen de norm, gebruikt bij toets *X*, handhaven bij toets *Y*.

---

[5] omdat zij geen rekening houden met het verschil in betrouwbaarheid tussen de toetsen en de deeltoetsen met gemeenschappelijke items



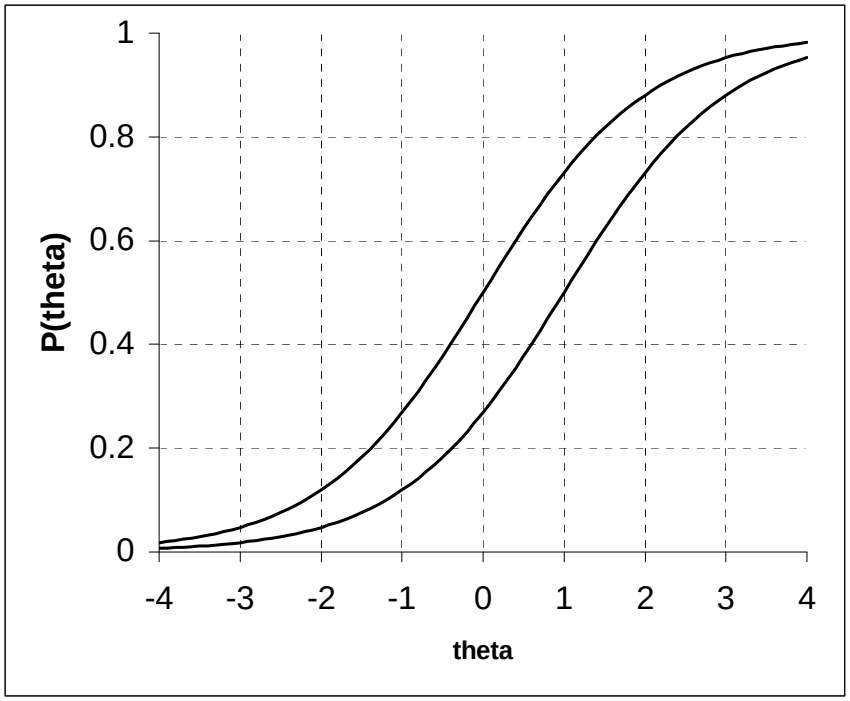
Figuur 4. Equipercentiel equivaleren; de cijfers geven de volgorde in het equivaleringsproces aan

In de praktijk wordt de methode uiteraard niet gebruikt. Als wij een test *Z* zouden hebben, dan zouden wij die willen meetellen voor het eindresultaat van een student. *Z* wordt dan een deeltoets van de tests (*X* + *Z*) en (*Y* + *Z*). De deeltoets met gemeenschappelijke items, *Z*, is dan hoe dan ook onbetrouwbaarder dan de totale toetsen.

Als *Z* een gemeenschappelijke deeltoets van een oude toets en een nieuwe toets is, dan is het nog steeds mogelijk om de oude en de nieuwe toets met elkaar te equivaleren al zullen wij de methode moeten aanpassen. Voorwaarde is wel dat *Z* op beide tijdstippen hetzelfde functioneert. De items uit *Z* mogen niet verouderd zijn wegens een aanpassing van het onderwijs. De items uit *Z* mogen evenmin na de eerste keer onder de studenten bekend zijn geworden.

Twee tests met een gemeenschappelijke deeltoets kunnen onder andere m.b.v. item-respons modellen worden geëquivaleerd. Grote toetsinstellingen zoals het CITO maken gebruik van dergelijke modellen. Om die reden zullen wij er kort aandacht aan besteden.

Wij veronderstellen dat aan de antwoorden één vaardigheid, beheersing van het vak, ten grondslag ligt. Wij noemen de mate van beheersing θ. De kans op een goed antwoord op een item neemt toe met een toename van θ.



Figuur 5. Kans op een goed antwoord voor twee items

In Figuur 5 wordt de kans op een goed antwoord als functie van de vaardigheid θ gegeven voor twee items. Om technische redenen hebben wij een representatie van de vaardigheidsschaal gekozen die loopt van min oneindig (geen beheersing) naar plus oneindig (perfecte beheersing). Het kansmodel uit de figuur is het zogenaamde Rasch model. Het Rasch model is een eenvoudig model: er wordt noch rekening gehouden met verschillen in item-discriminaties, noch met gokken.

Als alle items uit ($X + Z$) en ($Y + Z$) aan het Rasch model voldoen, dan kunnen wij de kanscurves (zie Figuur 5) van alle items op een gemeenschappelijke θ-schaal berekenen en vervolgens de relatie tussen de scores van ($X + Z$) en ($Y + Z$) bepalen. Voor informatie over equivaleren en over item-respons theorie, zie De Gruijter en Van der Kamp (2008).

16 Literatuur

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thornike (Red.), *Educational measurement* (2de editie). Washington, DC: American Psychological Association.

Bloom, B. S., Engelhart, M. D., Furst, E.J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. Handbook I: cognitive domain.* New York: McKay.

Coffman, W. E. (1971). Essay examinations. In: R. L. Thorndike (red.), *Educational Measurement.* Washington, DC: American Council on Education.

Cohen-Schotanus, J. (1994). *Effecten van curriculumveranderingen.* Groningen: proefschrift.

Cohen-Schotanus, J., Van der Vleuten, C. P. M., & Bender, W. (1996). Een betere cesuur bij tentamens, *Onderzoek van Onderwijs, 25*, 54-55.

De Groot, A. D., & Van Naerssen, R. F. (1969, 1973). *Studietoetsen.* Den Haag: Mouton.

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22*, 263-269.

De Gruijter, D. N. M. (1994). Kwaliteitseisen schriftelijke tentamens. *Onderzoek van Onderwijs, 23*, 10-12.

De Gruijter, D.N.M. (2000). De scoring van open vragen. In Heijnen, G., & Meeder, S. (Red.). *Toetsen en ICT in het hoger onderwijs, Stand van zaken en trends in Nederland, seminar, december 1999*, Utrecht: SURF.

**De Gruijter, D. N. M., & Van der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioral sciences.* Boca Raton, FL: Chapman & Hall/CRC.

Dousma, T., Hortsen, A., & Brants, J. (1997). *Tentamineren.* Groningen: Wolters-Noordhoff.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19*, 267-278.

Grosse, M. E., & Wright, B. D. (1985). Validity and reliability of true-false items. *Educational and Psychological Measurement, 45*, 1-13.

*Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 10, 159-170.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (red.), *Educational measurement*, (4th ed., pp. 433-470). Westport, CT: American Council on Education/Praeger.

Hofstee, W. K. B. (1977). Cesuurprobleem opgelost. *Onderzoek van Onderwijs, 6*, 6-7.

Messick, S. (1989). Validity. In R. L. Linn (Red.), *Educational Measurement, 3de editie*, New York: American Council on Education.

Millman, J. (1989). If at first you dont't succeed: setting passing scores when more than one attempt is permitted. *Educational Researcher, 18/8*, 5-9.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.

Van Berkel (1999). *Zicht op toetsen. Toetsconstructie in het hoger onderwijs.* Assen: Van Gorcum.

Van den Brink, W. P. (1977). Het verken-effect. *Tijdschrift voor Onderwijsresearch, 6*, 253-261.

Vos, P. (1992). Het ritme van het rooster. *Onderzoek van Onderwijs, 21/4*, 52-53.

*Wilcox, R. R. (1976). A note on the length and passing score of a mastery test. *Journal of Educational Statistics.* 1, 359-364.

Wijnen, W. H. F. W. (1971). *Onder of boven de maat; een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen.* Amsterdam: Swets & Zeitlinger.

*/** Vereist enige/redelijke wiskundige achtergrond.

Bijlage I: Enkele statistische begrippen

Wij hebben bij 20 studenten twee toetsen, toets *X* en toets *Y*, afgenomen. Beide toetsen bestaan uit 15 vragen. De scores van de studenten staan in Tabel I.1.

Tabel I.1. De scores van de studenten op toets *X* en toets *Y*.

student	score op *X*	score op *Y*	student	score op *X*	score op *Y*
1	5	4	11	12	13
2	4	9	12	14	14
3	12	9	13	5	6
4	11	11	14	6	6
5	10	9	15	11	7
6	10	7	16	14	14
7	9	5	17	13	13
8	9	8	18	14	8
9	10	4	19	13	10
10	12	11	20	10	14

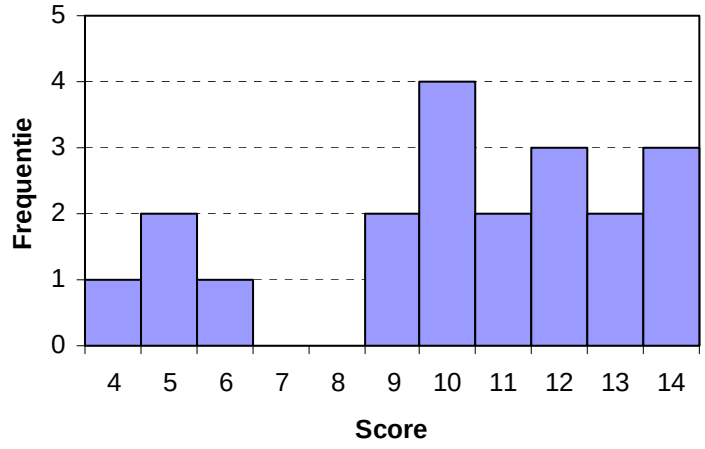
frequentieverdeling

In eerste instantie kijken wij alleen naar de scores op toets *X*. Wij kijken naar het aantal keren dat een bepaalde score voorkomt, de frequentie van deze score. De frequenties zetten wij in een frequentietabel, Tabel I.2. De frequenties zeggen niet zoveel: zij zijn afhankelijk van de omvang van de groep studenten die tentamen heeft gedaan. Daarom zetten wij in Tabel I.2
naast de frequentie het percentage studenten dat een bepaalde score heeft behaald. In de kolom daarnaast staat het cumulatief percentage: het percentage studenten met de gegeven score of een lagere score. Uit de tabel kunnen wij aflezen dat 20 procent van de studenten een score van 6 of minder heeft behaald. Als de score 10 de laagste voldoende zou zijn, dan zou 30 procent van de studenten zijn gezakt. In de laatste kolom staat de percentielscore. De percentielscore van een student geeft aan hoeveel procent van de studenten een lagere score heeft gehaald. Voor een student met een score van 10 zou de percentielscore 30 zijn: 30 procent van de studenten heeft een lagere score. Bij de berekening van de percentielscore wordt echter dikwijls gedaan alsof wij het percentage in een bepaalde categorie kunnen ‘smeren’ over het categorie-interval. Van het percentage studenten dat de score 10 heeft behaald, wordt de helft geacht te vallen in het interval 9.5 – 10, de andere helft wordt geacht te vallen in het interval 10 –10.5. Als wij zo te werk gaan is de percentielscore van een student met score 10 gelijk aan het cumulatief percentage behorend bij score 9 (30 %) plus de helft van het percentage behorend bij score 10 (10 %): 40 procent.

Tabel I.2. De verdeling van de scores op toets *X*.

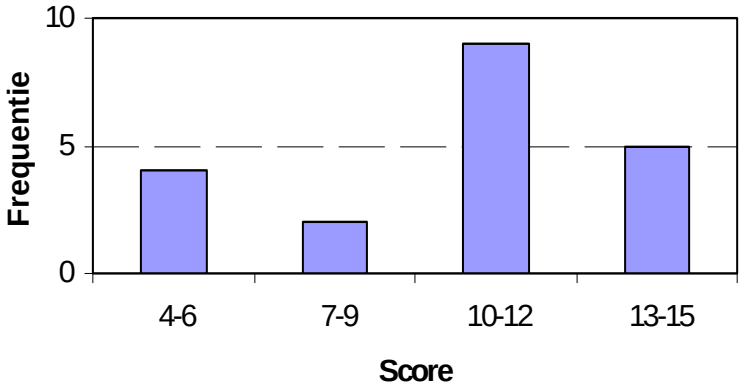
Score	Frequentie	Percentage	Cumulatief percentage	Percentielscore
4	1	5.0	5.0	2.5
5	2	10.0	15.0	10.0
6	1	5.0	20.0	17.5
9	2	10.0	30.0	25.0
10	4	20.0	50.0	40.0
11	2	10.0	60.0	55.0
12	3	15.0	75.0	67.5
13	2	10.0	85.0	80.0
14	3	15.0	100.0	92.5

De verdeling van de scores op toets *X* kan ook grafisch worden weergegeven. Dat wordt in Figuur I.1 gedaan. Bij de constructie van deze figuur is gebruik gemaakt van een rekenblad (Excel); dat geldt overigens voor alle figuren en berekeningen in deze bijlage.



Figuur I.1. De verdeling van de scores op toets *X*

Wellicht is deze figuur wat druk. Als er veel verschillende categorieën in de figuur voorkomen, is het handiger om categorieën samen te voegen, zoals in Figuur I.2.



Figuur I.2. De verdeling van de scores op toets *X*; gegroepeerd

Gemiddelde, variantie en standaardafwijking

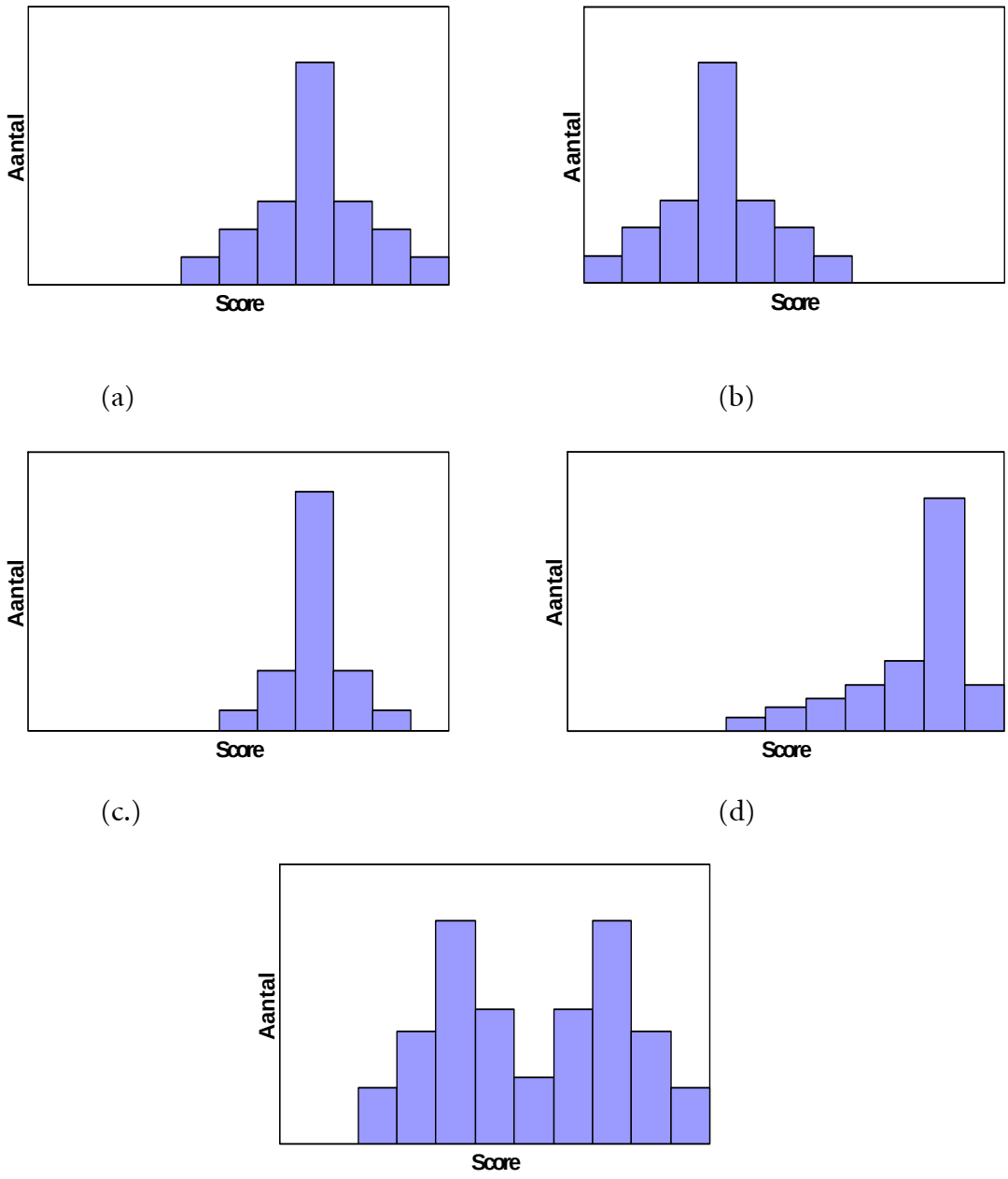
Relatief hoge scores komen het meeste voor. De verdeling ligt op een schaal van 0 (de minimale score) tot 15 (de maximale score) aan de rechterkant. De positie van de verdeling wordt op beknopte wijze vastgelegd in een centrummaat. Wij behandelen het gemiddelde (andere centrummaten zijn de *modus* en de *mediaan*).

Het gemiddelde van de scores van toets *X* is gelijk aan de som van de scores, gedeeld door het aantal scores. Het gemiddelde geven wij aan met $\bar{x}$. Het gemiddelde wordt dus berekend als:

$$\bar{x} = \sum_{i=1}^{N} x_i / N \,, \qquad \text{(I.1)}$$

waarbij Σ het symbool voor sommeren is, x_i de score is van de *i*-de student en *N* het totaal aantal studenten. Het gemiddelde voor toets *X* is gelijk aan 10.20.

Verschillende verdelingen kunnen qua gemiddelde verschillen. In Figuur I.3a en I.3b staan twee verdelingen die alleen qua gemiddelde verschillen. In Figuur I.3b staat een verdeling met een lager gemiddelde dan in Figuur I.3a.



Figuur I.3. Verschillende verdelingsvormen

Verdelingen kunnen op nog meer manieren verschillen. In Figuur I.3c is de spreiding van de scores rond het gemiddelde kleiner dan in Figuur I.3a. Figuur I.3d geeft een scheve verdeling en Figuur I.3e is meertoppig. Wij zullen hier verder alleen nog het verschil in spreiding aan de orde stellen. daarvoor gebruiken wij de begrippen *variantie* en *standaardafwijking*.

Wij berekenen eerst de variantie, in formule:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}. \qquad \text{(I.2)}$$

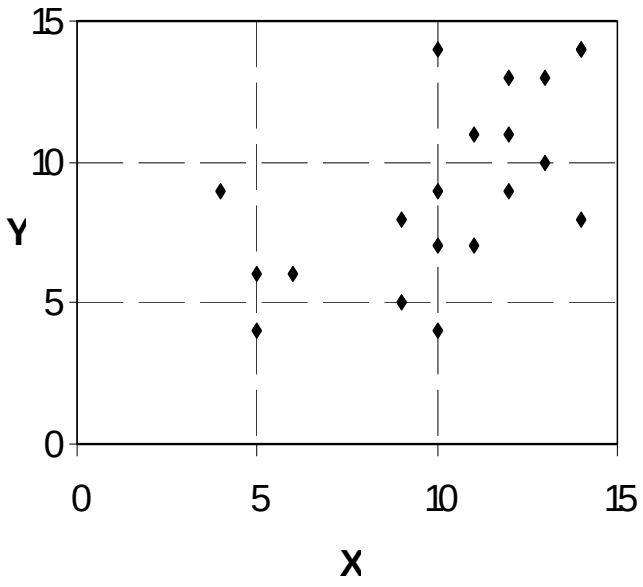
In de formule worden de N afwijkingen van het gemiddelde genomen. De afwijkingen worden gekwadrateerd. Van de gekwadrateerde afwijkingen wordt vervolgens een soort gemiddelde genomen. Op statistische gronden is deling van de noemer door $N - 1$ veelal te prefereren boven deling door N, maar voor de eenvoud zullen wij hier (I.2) gebruiken (in het rekenblad zijn overigens beide berekeningswijzen mogelijk). De variantie van de toetsscores X uit Tabel I.1 is 9.16.

Een belangrijke maat voor de spreiding van gegevens is de *standaardafwijking s*. De standaardafwijking is gelijk aan de (vierkants)wortel uit de variantie. De standaardafwijking van de scores van toets X uit Tabel I.1, s_X, is gelijk aan 3.03.

Wat gebeurt er met de variantie en de standaardafwijking als wij met proporties goede antwoorden in plaats van met scores x zouden willen werken? De scores x worden dan door 15 gedeeld. De variantie wordt een factor (15×15) kleiner, en de standaardafwijking wordt een factor 15 kleiner.

Correlatie

Laten wij nu de twee toetsen X en Y in hun samenhang bekijken. Wij kunnen de scores op Y afzetten tegen die van X in een puntenwolk. Dat is gedaan in Figuur I.4.



Figuur I.4. Puntenwolk van de scores op X en Y

De combinatie van scores 14, 14 komt tweemaal voor. Dat detail is uit deze figuur niet op te maken. Wel is duidelijk te zien dat de scores op X en Y enige samenhang vertonen. Hogere scores op X gaan gemiddeld samen met hogere scores op Y. Een

maat voor de lineaire samenhang tussen twee variabelen *X* en *Y* wordt door de correlatie weergegeven. Daarvoor introduceren wij eerst de gestandaardiseerde of z-score. Voor toets *X* worden de z-scores gegeven door:

$$z_{x_i} = \frac{x_i - \bar{x}}{s_X}, \tag{I.3}$$

en voor Y verkrijgt men op dezelfde wijze z-scores. De z-scores zijn schaalonafhankelijk: telt men bij alle scores een constante op of vermenigvuldigt men de scores met een constante, dan blijft de z-score van een persoon gelijk. De correlatie tussen *X* en *Y*, r_{xy}, is met z-scores eenvoudig te schrijven als:

$$r_{xy} = \frac{\sum_{i=1}^{N} z_{x_i} z_{y_i}}{N}. \tag{I.4}$$

De correlatie tussen *X* en *Y* uit Tabel I.1 is gelijk aan 0.61.

De correlatie heeft als maat van samenhang de volgende eigenschappen:

- De correlatie is schaalonafhankelijk: als men alle getallen *x* met een constante vermenigvuldigt of er een constante bij optelt, dan verandert de waarde van de correlatie niet. Als *x* op temperaturen zou slaan, zou het dus niet uitmaken of wij in graden Celsius of in graden Fahrenheit gemeten hebben.
- Het maximum van de correlatie is 1. Het maximum wordt bereikt als het verband tussen X en Y volstrekt lineair is: $x = ay + b$, met $a > 0$. Het minimum is –1 als $x=ay+b$, met $a < 0$. Als er geen *lineair* verband is, is de correlatie gelijk aan 0.
- De correlatie kan niet bepaald worden als de variantie van één van de variabelen 0 is.

Als men de correlatie tussen *X* en *Y* kent, dan kan men de waarde van *Y* 'voorspellen' uit de waarde van *X* via de zogenaamde regressielijn voor de regressie van *Y* op *X*:

$$\hat{y} = s_y r_{xy} (x - \bar{x}) / s_x + \bar{y}, \tag{I.5}$$

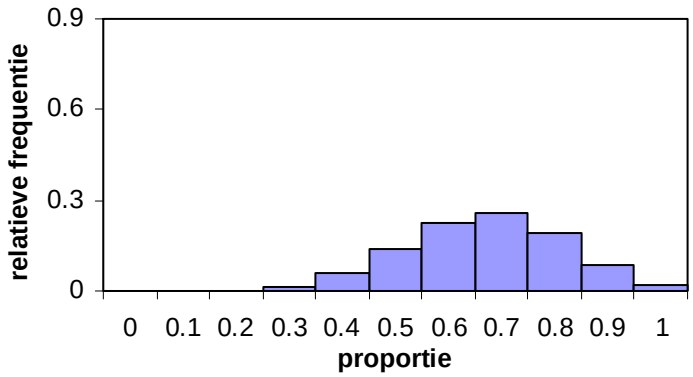
waarbij het 'dakje' boven de *y* aangeeft dat wij met een schatting te maken hebben.

inferentiële statistiek; een voorbeeld

Het gemiddelde en de variantie van een verdeling worden dikwijls niet alleen gebruikt om een verdeling te beschrijven. Veelal willen wij uit de gegevens gevolgtrekkingen maken. Wij hebben bijvoorbeeld een tentamenvraag met een *p*-waarde, de proportie goede antwoorden, gelijk aan 0.80 en constateren dat de vraag gemakkelijk is. Daarbij bedoelen wij meer te zeggen dan dat de vraag binnen

de groep studenten die tentamen heeft gedaan, goed is gemaakt. Wij veronderstellen dat de vraag ook bij soortgelijke groepen studenten in de toekomst relatief goed zal worden beantwoord. Wij zullen de gevolgtrekking over de moeilijkheidsgraad van de vraag niet maken als slechts vier studenten aan het tentamen deelnamen. In een andere groep studenten zou de *p*-waarde gemakkelijk veel lager kunnen uitvallen. Hoe groter de groep studenten is, des te meer vertrouwen we hebben op de juistheid van een generalisatie van het resultaat naar soortgelijke groepen studenten.

Wij diepen het voorbeeld van een proportie wat verder uit. Dat doen wij aan de hand van een heel doorzichtig kansmechanisme: het werpen van dobbelstenen. Laten wij een worp van 3 of hoger een treffer noemen. De kans op een treffer,π[6], is dan 2/3, ten minste als wij met een zuivere dobbelsteen werpen. Laten wij nu 10 worpen achter elkaar doen. De kansen op 0 treffers uit 10 worpen, 1 treffer uit 10 worpen, etc. staan in Figuur I.5. de kans op een proportie van 0.8 treffers is gelijk aan 0.20, de kans op een proportie van 0.8 of hoger is 0.30. Laten wij nu weer teruggaan naar de *p*-waarde bij een tentamen. Van het voorbeeld van de dobbelsteen kunnen wij leren dat als de proportie correct in de relevante populatie van studenten 2/3 is, er in een groep van 10 willekeurige studenten, een aselecte steekproef van 10 studenten uit de populatie, een grote kans is dat de geobserveerde *p*-waarde 0.80 of hoger is[7].



Figuur 1.5. De kansverdeling voor de proportie treffers bij 10 metingen

Het gemiddelde van de verdeling van proporties *p* in Figuur 1.5 is gelijk aan de kans op een treffer π. De standaardafwijking is:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}}, \qquad \text{(I.6)}$$

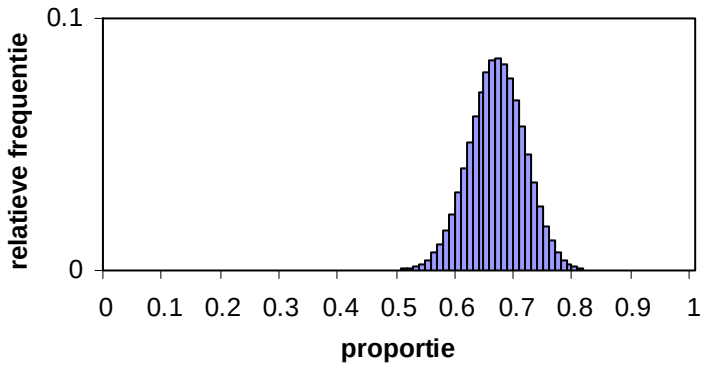
waarbij *N* het aantal worpen (hier 10) is. De standaardafwijking van de proporties in Figuur I.5 is 0.15. De verdeling in Figuur I.5 is niet symmetrisch: de staart aan de linkerkant is groter dan die aan de rechterkant.

---

6 Wij kiezen een Griekse letter om een onderscheid te maken met geobserveerde proporties.

7 Deze uitspraak mogen wij doen als wij de relevante populatie als heel groot mogen beschouwen, veel groter dan het aantal studenten bij het tentamen.

Wat gebeurt er nu met de kansverdeling als wij de proportie treffers berekenen bij 100 worpen met de dobbelsteen.? Het effect is te zien in Figuur I.6.



Figuur I.6. De kansverdeling voor de proportie treffers bij 100 metingen

Het gemiddelde van de steekproefverdeling van proporties *p* is gelijk aan de kans op een treffer π. De standaardafwijking is bij steekproeven van 100 metingen kleiner dan bij 10 metingen (zie Formule(I.6)). In de figuur zien wij dat de verdeling meer gecentreerd is rond het gemiddelde. De verdeling oogt fraaier. De verdeling begint te neigen naar de klokvorm van de zogenaamde normale verdeling. Daar gaan wij gebruik van maken.

In de normale verdeling ligt 95 procent van de verdeling tussen het gemiddelde van de verdeling min 1.96 maal de standaardafwijking en het gemiddelde plus 1.96 maal de standaardafwijking. Wij kunnen dus zeggen dat bij benadering 95 procent van de verdeling in Figuur I.6 ligt tussen 0.57 en 0.76.

Daar hebben wij nog niet zoveel aan. Het wordt pas interessant als wij het gemiddelde van de verdeling, π, niet kennen, maar wel een proportie *p* hebben geobserveerd. Kunnen wij op grond van de geobserveerde proportie p een uitspraak over π doen? Het antwoord is ja. Als wij de standaardafwijking σ van de steekproefverdeling kennen, dan mogen wij de uitspraak doen dat de de kans π mat 95-procent zekerheid ligt in het interval van $p - 1.96\sigma$ tot $p+1.96\sigma$. Het desbetreffende interval wordt een *betrouwbaarheidsinterval* genoemd. Ook deze uitkomst is nogal theoretisch: wij kennen σ, de standaardmeetfout van het gemiddelde, niet. Die schatten wij met[8]:

$$s_p = \sqrt{\frac{p(1-p)}{N}}\,. \qquad \text{(I.7)}$$

Zo kunnen wij dus een uitspraak doen over de moeilijkheidsgraad van een item op basis van de empirische *p*-waarde. Laten wij er weer vanuit gaan dat wij een *p*-waarde gevonden hebben van .80, maar nu bij een aselecte steekproef van 100 studenten. Dan kunnen wij ervan uitgaan dat de proportie correct in de populatie, π, naar alle waarschijnlijkheid ligt tussen .72 en .88.

---

8 Ook dat mogen wij alleen doen als de proportie op een redelijk aantal metingen is gebaseerd omdat de standaardmeetfout in het relevante gebied niet teveel met π mag variëren.

Het model dat de kans op een bepaalde uitslag bij het werpen van dobbelstenen geeft, heet het binomiale model. In de testtheorie wordt het model ook gebruikt als foutenmodel bij een bepaald soort toetsingen. Veronderstel dat wij een grote voorraad vragen hebben. De beheersing van een student definiëren wij als de proportie vragen π uit de grote voorraad die deze student goed kan beantwoorden. Wij nemen willekeurig n vragen uit de vragenvoorraad. De kans op 0, 1, . . . ,n goede antwoorden wordt dan gegeven door de binomiale verdeling gegeven de waarde van π.

Bijlage II: Een inleiding in de klassieke testtheorie

Als wij een student twee vergelijkbare toetsen laten maken, dan is de totaalscore op die toetsen naar alle waarschijnlijkheid verschillend. Onze inschatting van het beheersingsniveau van de student hangt dus van de toevallige toets af. De uitkomst is niet precies: wij hebben met meetfouten te maken. Als wij de grootte van de meetfout zouden kennen en die van de geobserveerde score zouden aftrekken, dan houden wij de score zonder meetfout over, de ware score. In de klassieke testtheorie worden uitspraken gedaan over de invloed van meetfouten op de scores binnen populaties of groepen van personen. Het blijkt mogelijk om verschillende soorten meetfouten te onderscheiden, afhankelijk van het doel dat met de metingen wordt nagestreefd. Een wijziging van de definitie van de meetfout gaat uiteraard gepaard met een wijziging in de definitie van de ware score. Een voorbeeld. Veronderstel dat wij een test hebben voor een stabiele eigenschap. Dan mogen wij de variatie van scores op twee verschillende momenten als ruis beschouwen. Bij de toetsing van beheersing van de bestudeerde stof in het onderwijs gaan wij er niet vanuit dat wij met een stabiele eigenschap van een student te maken hebben. Variatie van de scores van een student in de tijd is dan voor een deel te beschouwen als een verandering van de ware score van deze student. Wij zullen terugkomen op de definitie van ware score en meetfout bij de bespreking van de schatting van de betrouwbaarheid van een toets.

Laten wij eerst van één student, student p, uitgaan. De geobserveerde score van deze student is x_p. Het idee is dat wij deze score kunnen schrijven als de som van een ware score τ_p en een meetfout e_p: (in het engels 'true score' en 'measurement error'):

$$x_p = \tau_p + e_p. \qquad \text{(II.1)}$$

In plaats van de score x_p hadden wij mogelijkerwijze een andere score verkregen. Deze niet waargenomen scores geven wij met een hoofdletter X aan, ter onderscheiding van de geobserveerde score. In het algemeen kunnen wij dus schrijven:

$$X_p = \tau_p + E_p. \qquad \text{(II.2)}$$

Veronderstel nu dat wij in staat zijn om herhaald metingen X_p onafhankelijk van elkaar te verrichten. Wij kunnen de ware score zien als het gemiddelde van veel onafhankelijke metingen. Het gemiddelde van de meetfouten is dus gelijk aan 0 als wij maar genoeg metingen hebben verricht.[9]

Nu nemen wij een populatie van studenten. In de populatie is de gemiddelde meetfout ook 0 (bij voldoende metingen!) en de correlatie tussen de toevallige meetfouten en de ware scores is eveneens gelijk aan 0. Wij kunnen de variantie van de geobserveerde scores daarom schrijven als:

$$\sigma^2_X = \sigma^2_T + \sigma^2_E. \qquad \text{(II.3)}$$

---

9 de ware score en meetfout krijgen in de testtheorie een nauwkeuriger definitie dan de losse omschrijving die wij hier gebruiken

De geobserveerde variantie is gelijk aan de variantie van de ware scores plus de variantie van de meetfouten.

De *betrouwbaarheid* van een toets wordt gedefinieerd als de verhouding tussen de ware-score variantie en de geobserveerde-score variantie:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}. \qquad \text{(II.4)}$$

Op het symbool $\rho_{XX'}$ gaan wij hieronder nader in. Laten wij eerst nagaan wat betrouwbaarheid inhoudt. De betrouwbaarheid is minimaal gelijk aan 0. Dat is het geval als de variantie van de ware scores gelijk is aan 0. Wij hebben dan alleeen met ruis te maken. De betrouwbaarheid is maximaal gelijk aan 1, als de meetfoutvariantie gelijk is aan 0. De betrouwbaarheids-coëfficiënt geeft aan in hoeverre geobserveerde verschillen verschillen tussen ware scores reflecteren. De definitie geeft aan dat de betrouwbaarheidscoëfficiënt populatieafhankelijk is. In een populatie waarin de ware verschillen tussen personen relatief klein zijn, valt de betrouwbaarheidscoëfficiënt lager uit dan in een populatie met een grote variatie.In veel toepassingen is het maken van onderscheid tussen personen belangrijk en behoort het meetinstrument betrouwbaar te zijn. Bij het toetsen van beheersing van de stof in het onderwijs gaat het niet primair om het maken van onderscheid tussen studenten. Er kunnen zich situaties voordoen waarin de toets voldoet ondanks het feit dat de schatting van de betrouwbaarheid in de desbetreffende groep studenten laag is.

Parallelle tests en betrouwbaarheid

Onder parallelle tests worden tests verstaan die volstrekt uitwisselbaar zijn. Parallelle tests hebben dezelfde ware scores en voor elke persoon dezelfde meetfoutenvariantie. Voor twee parallelle tests X en X' geldt dus onder andere:

$\mu_X = \mu_{X'}$, de gemiddelden van X en X' zijn in de populatie gelijk,

$\sigma^2_E = \sigma^2_{E'}$

$\sigma^2_T = \sigma^2_{T'}$

en

$\sigma^2_X = \sigma^2_{X'}$.

Men kan afleiden dat de *correlatie tussen twee parallelle tests X en X', $\rho_{XX'}$ gelijk is aan de betrouwbaarheid.*

Daarmee is een eerste methode gevonden om de betrouwbaarheid van een toets X te schatten. Wij construeren een tweede parallelle toets X', nemen beide toetsen af en correleren de uitkomsten. Aan deze methode kleven toch wel nadelen. Als wij twee toetsen X en X' afnemen, zullen de studenten van ons eisen dat wij uitslagen

berekenen op basis van de score $X + X'$. De betrouwbaarheid van deze tweemaal zo lange toets is hoger dan de betrouwbaarheid van de oorspronkelijke toetsen. Wij komen nog terug op de schatting van de betrouwbaarheid van verlengde toetsen. Andere bezwaren zijn dat de definitie van parallelliteit niet eenduidig is – wij kunnen op verschillende manieren parallelle toetsen maken – en dat wij niet weten of de toetsen X en X' inderdaad parallel zijn. Wij kunnen hoogstens nagaan of de toetsen hetzelfde gemiddelde en dezelfde variantie hebben.

Testverlenging

Wij hebben twee parallelle toetsen X en X' en wij willen de betrouwbaarheid schatten van de toets $X + X'$. De betrouwbaarheid van de toets na toetsverlenging met de factor 2 schrijven wij als $\rho_{X(2)X'(2)}$. De waarde van de betrouwbaarheid van de verlengde toets is:

$$\rho_{X(2)X(2)'} = \frac{2\rho_{XX'}}{1+\rho_{XX'}}. \qquad \text{(II.5)}$$

Formule (II.5) is de Spearman-Brown formule voor testverlenging. Met behulp van deze formule is de eerste schatting van de betrouwbaarheid mogelijk. Als wij van een toets de betrouwbaarheid willen weten, dan delen wij de test in twee gelijke helften, correleren de twee resulterende deeltoetsen en gebruiken Formule (II.5) voor een schatting van de betrouwbaarheid van de totale toets. Deze procedure is één van de mogelijke 'split-half' schattingsmethoden.

Van Formule (II.5) bestaat een generalisatie naar testverlenging met een factor k:

$$\rho_{X(k)X(k)'} = \frac{k\rho_{XX'}}{1+(k-1)\rho_{XX'}}. \qquad \text{(II.6)}$$

In de praktijk wordt veelal coëfficiënt α voor de schatting van de betrouwbaarheid gebruikt:

$$\alpha \equiv \left(\frac{k}{k-1}\right)\left(1-\frac{\sum_{i=1}^{k} s_i^2}{s_X^2}\right), \qquad \text{(II.7)}$$

waarbij k het aantal items in de toets is, s_i^2 de variantie van item i, en s_X^2 de variantie van de totaalscores op de toets. Coëfficiënt α is een onderschatting van de betrouwbaarheid; bij niet al te korte toetsen valt de mate van onderschatting meestal mee[10]. Een betere ondergrens voor de betrouwbaarheid is λ_2, dat in verschillende softwarepakketten beschikbaar is. Voor items die goed-fout worden gescoord, is coëfficiënt α te herschrijven tot de KR20, de Kuder-Richardson

[10] De coëfficiënt kan zelfs negatief zijn, de betrouwbaarheid zelf is nooit negatief.

formule 20 naar de auteurs van de formule Kuder en Richardson. Een vereenvoudigde vorm van de KR20 is:

$$\mathrm{KR}21 \equiv \left(\frac{k}{k-1}\right)\left(1-\frac{k\bar{p}(1-\bar{p})}{s_X^2}\right), \tag{II.8}$$

waar $\bar{p}$ de gemiddelde *p*-waarde is. De waarde van KR21 is lager dan de waarde van KR20 tenzij alle items even moeilijk zijn. Nu alle berekeningen met de computer gebeuren, is de formule niet meer van belang voor het schatten van de betrouwbaarheid van een toets. De formule blijkt wel relevant als elke student een andere steekproef van items uit een grote itemvoorraad krijgt.

Soms bestaat de toetsing uit meer dan één onderdeel. Dan kan de betrouwbaarheid van de totaalscore worden berekend met behulp van de formule:

$$r_{XX'} = \frac{\sum_{i=1}^{q} r_{ii'} s_{Y_i}^2 + \sum_{i=1}^{q}\sum_{j\neq i}^{q} s_{Y_i Y_j}}{s_X^2}, \tag{II.9}$$

waarbij er q onderdelen Y_1, Y_2, . . . Y_q zijn met X als som en r_{ii} de betrouwbaarheidsschatting voor onderdeel Y_i is. Een generalisatie naar de situatie waarbij de afzonderlijke onderdelen met een verschillend gewicht meetellen, is eenvoudig te maken.

De standaardmeetfout

De standaardafwijking van de meetfouten wordt de standaardmeetfout genoemd. Als wij de betrouwbaarheid hebben geschat, is de standaardmeetfout te berekenen via:

$$s_E = s_X \sqrt{1-r_{xx'}}\,. \tag{II.10}$$

De grootte van de standaardmeetfout geeft aan hoe onnauwkeurig toetsscores zijn. De standaardmeetfout wordt wel gebruikt om een betrouwbaarheidsinterval voor een ware score te maken (zie Bijlage I). Het gebruik van de standaardmeetfout daarvoor is helaas niet onproblematisch. Bij personen met heel hoge ware scores is de variantie van de meetfouten kleiner dan bij personen met wat lagere ware scores. De geobserveerde scores kunnen immers niet hoger worden dan de maximale toetsscore. Bij personen met een hoge ware score is er sprake van een 'plafondeffect'. De formule (II.9) geeft slechts een schatting van de gemiddelde waarde over alle personen die de toets hebben afgelegd.

De schatting van de ware score

Wij zijn in de geobserveerde score van een student geïnteresseerd omdat deze een indicatie geeft van de prestaties die wij van deze persoon mogen verwachten. Wij zijn dus eigenlijk in de ware score geïnteresseerd. De geobserveerde score geeft een

schatting van de ware score. In toepassingen gebruiken wij de geobserveerde score alsof deze de ware score is. Als de standaardmeetfout relatief gering is, is dat geen probleem.

De geobserveerde score is niet de enige schatting van de ware score en in statistisch opzicht ook niet de beste. In Bijlage I zijn wij regressieformule (I.5) tegengekomen met behulp waarvan wij een variabele kunnen schatten op basis van een andere variabele. Wij kunnen deze formule gebruiken om de ware score te schatten uit de geobserveerde score. In dit geval kan de formule geschreven worden als de zogenaamde Kelley-formule:

$$\hat{\tau} = r_{XX'}(x - \bar{x}) + \bar{x}\,. \qquad \text{(II.11)}$$

Als de betrouwbaarheid gelijk is aan 0 geeft de Kelley-schatter voor alle personen dezelfde schatting van de ware score, de gemiddelde score. Dat is logisch. Als de betrouwbaarheid gelijk is aan 0, dan weten wij immers dat alle geobserveerde variatie meetfouten betreft.

Toch geeft het gebruik van de Kelley-schatter ook problemen, zoals

- Een persoon kan lid zijn van meer groepen. Van welke groep gebruiken wij gegevens t.b.v. de formule?
- Het is niet te verkopen als men personen met dezelfde score die uit verschillende populaties komen, verschillend behandelt.
- De betrouwbaarheid moet worden geschat en dat gebeurt met een wellicht onnauwkeurige onderschatting.

auteursregister

zakenregister