Bulletin Medisch Onderwijs, 1994, 13, 37-43

Statistiek en meten: wat moet je daarover weten?

Onder redactie van Diana Dolmans, Cees van der Vleuten, Albert Scherpbier en Ineke Wolfhagen

Bij het lezen van publikaties over onderzoek van onderwijs worden lezers regelmatig geconfronteerd met allerlei statistische begrippen waarvan de betekenis voor velen niet geheel duidelijk is. Daarom heeft de redactie besloten een reeks artikelen te publiceren waarin belangrijke onderwerpen op het gebied van statistiek en meten aan de orde worden gesteld. De redactie van deze reeks wordt gevormd door twee gastredacteuren (Diana Dolmans en Cees van der Vleuten) en twee leden van de BMO-redactie (Albert Scherpbier en Ineke Wolfhagen).
De keuze van onderwerpen is gebaseerd op veel gebruikte statistische technieken en begrippen. De nadruk ligt op de betekenis en interpretatie hiervan. Op deze manier wordt getracht een bijdrage te leveren aan de verdere professionalisering van docenten binnen het medisch onderwijs. In het vijfde artikel van deze reeks staat een aantal begrippen uit de item-analyse centraal.

Item-analyse

W. Bender

Termen die aan bod komen:
scorematrix, ruwe score, p-waarde, a-waarde, item-totaal correlatie (Rit), item-rest correlatie (Rir), betrouwbaarheid (KR-20), Signal-Noise-ratio, standaardmeetfout (smf)

"Als u mij wilt toestaan om voor de n'de keer een definitie van statistiek te geven, dan wil ik statistiek definiëren als de kunst van het nuttig verliezen van informatie."[1] Item-analyse is een toepassing van deze verlieskunde op het gebied van testen of toetsen of tentamens. De complete informatie betreffende een afgelegde toets wordt weergegeven door de *scorematrix*, dat is een tabel waarin de rijen worden gevormd door personen en de kolommen door vragen. Een kruispunt (cel) bevat een code die het antwoord van de betreffende persoon op de betreffende vraag weergeeft. Om dit nader toe te lichten zal ik mij beperken tot geprecodeerde vragen, dat wil zeggen vragen waarop antwoord kan worden gegeven door te kiezen uit twee of meer alternatieven. De volksmond noemt dit veelal multiple-choice-vragen, hetgeen minder gewenst is, omdat men daarbij ook een vraag met twee alternatieven als een multipele keuze betitelt.

Stel dat een toets bestaat uit een aantal vierkeuzevragen. Een scorematrix zou er uit kunnen zien zoals weergegeven in tabel 1. Afgebeeld zijn de antwoorden van acht personen op zes vragen (een geprecodeerde vraag wordt vaak aangeduid als 'item' - naar keuze op z'n Nederlands of op z'n Engels uitgesproken). Bij een vierkeuzevraag zijn vier antwoorden mogelijk en de cijfers in de cellen corresponderen met de feitelijk gekozen antwoorden. Dus, persoon 1 heeft het vierde alternatief op vraag 1 als het goede antwoord gekozen. De andere personen hebben een andere keuze gemaakt. Persoon 7 heeft vraag 1 overgeslagen (e causa ignota). Over vraag 4 en vraag 5 zijn alle personen het roerend eens: zij kiezen als het goede antwoord respectievelijk het tweede en het vierde alternatief.

Tabel 1. Een voorbeeld van een ruwe scorematrix

	item					
persoon	1	2	3	4	5	6
1	4	3	2	2	4	2
2	3	3	3	2	4	2
3	1	1	1	2	4	2
4	2	2	2	2	4	2
5	2	4	4	2	4	4
6	1	2	3	2	4	2
7		1	2	2	4	4
8	1	4	3	2	4	1

Tabel 2. De antwoordsleutel

	item					
	1	2	3	4	5	6
sleutel	3	4	2	2	4	1

Of de gekozen antwoorden ook de goede antwoorden zijn, hangt uiteraard af van de antwoordsleutel. Als bij vraag 1 het derde alternatief het goede antwoord is, dan scoort alleen persoon 2 een punt op die vraag. Als bij vraag 4 het eerste alternatief als sleutel wordt opgegeven, dan zal niemand een punt op vraag 4 hebben verworven. Om te weten wat alle personen op de toets hebben gescoord, en om uit te maken hoeveel goede antwoorden voor elke vraag gegeven zijn, moet iedere rij van de scorematrix worden vergeleken met de sleutel. De sleutel is weergegeven in tabel 2. Bij vraag 1 is het derde alternatief het goede antwoord, bij vraag 2 het vierde alternatief, en zo vervolgens.

Wat is item-analyse?

In 1961 verscheen in de Verenigde Staten Multiple-choice examinations in medicine.[2] Dit nog steeds zeer lezenswaardige boekje, gepubliceerd vanuit de National Board of Medical Examiners, bevat de volgende inleiding op het onderwerp item-analyse. "After a multiple-choice test has been used for a large number of examinees, a vast amount of data becomes available for statistical analysis. The scores for individuals on the test as a whole, on subdivisions of the test, on the individual items, as well as the mean scores made by groups of individuals, may be studied separately and in relation to each other. Information may be obtained for the internal purposes of verifying the precision of the test as a measuring instrument and determining the extent to which it achieves the objectives of the examiner. Analyses may also be made to serve the external purpose of describing in considerable detail the performance of groups of examinees such as medical school classes or groups of physicians".

Item-analyse is dus de statistische bewerking van de scorematrix, waarmee de volgende waarden worden verkregen:
a. de score per persoon;
b. de p-waarde en de a-waarde;
c. de item-totaal correlatie (Rit);
d. de betrouwbaarheid (KR-20);
e. de standaarddeviatie (sd);
f. de standaardmeetfout (smf);
Hiermee zijn de belangrijkste elementen van de item-analyse wel genoemd, maar nog niet uitgelegd. Dat laatste kan het beste aan de hand van een concrete statistische bewerking van een toets (meestal aangeduid met het niet-bestaande woord uitdraai). Het voorbeeld betreft het item-analyseprogramma zoals dat aan de Rijksuniversiteit in Groningen door het Centrum Onderzoek Wetenschappelijk Onderwijs Groningen (COWOG) bij tentamina wordt gebruikt. De uitdraai wordt hier in twee gedeelten gepresenteerd: tabel 3 vermeldt de gegevens per vraag en tabel 4 bevat de frequentieverdeling van de *ruwe scores*. Het voorblad van de uitdraai, dat hier niet wordt afgebeeld, is een soort colofon. Het bevat gegevens als de naam van de toets, de datum van afname, het aantal vragen, het aantal kandidaten, plus nog enkele technische details die hier minder terzake zijn.

Tabel 3. Voorbeeld van een gedeelte van een uitdraai van de item-analyse

Vraagnummer	Sleutel	p-waarde	Item-test correlatie	Geen antwoord	1	2
1	2	0.954	–0.028	0.0	4.6	95.4*
2	2	0.934	0.200	0.0	6.6	93.4*
3	1	0.908	0.085	0.0	90.8*	9.2
4	1	0.857	0.121	0.0	85.7*	14.3
5	2	0.413	0.030	0.0	58.7	41.3*
6	1	0.883	0.372	0.0	88.3*	11.7
7	2	0.776	0.460	0.0	22.4	77.6*
8	2	0.321	–0.068	0.0	67.9	32.1*
9	2	0.934	0.235	0.0	6.6	93.4*
99	2	0.878	0.116	0.0	12.2	87.8*
100	1	0.985	0.147	0.0	98.5*	1.5
101	1	0.995	0.040	0.0	99.5*	0.5
102	1	0.872	0.154	0.0	87.2*	12.8
103	2	0.944	0.082	1.5	4.1	94.4*
104	2	0.954	0.134	1.5	3.1	95.4*
105	1	0.699	0.027	1.5	69.9*	28.6
106	1	0.827	0.214	1.5	82.7*	15.8
107	2	0.429	0.150	1.5	55.6	42.9*
108	2	0.184	0.182	2.0	79.6	18.4*
109	1	0.684	0.363	2.0	68.4*	29.6

Gemiddelde	77.240
Getelde vragen	100
Standaarddeviatie personen	5.872
Gemiddelde moeilijkheidsgraad	0.772
Correlatie significant op 5 procent niveau	0.139
KR-20	0.622
Signal-Noise-ratio	1.643
Standaardmeetfout	3.612

Maar werkelijk opwindend wordt de uitdraai pas na het voorblad (tabel 3).

Afgebeeld is een gedeelte van een in werkelijkheid twee pagina's omvattende tabel met een groot aantal gegevens per item. In de meest linkse kolom staan de vraagnummers, met daarnaast in de tweede kolom de antwoordsleutel. Het gaat om juist / onjuist vragen, en het goede antwoord op vraag 1 is 'onjuist' (hier gecodeerd als 2). Er is dus geen vraagtekenoptie in dit voorbeeld.

De derde kolom is de *p-waarde*, ook wel moeilijkheidsgraad genoemd. Het betreft de proportie personen die de betreffende vraag goed hebben beantwoord. Zo heeft item 1 een p-waarde van 0.954, hetgeen overeenkomt met 187 (van de 196) kandidaten die het goede antwoord 'onjuist' hebben gegeven. Klaarblijkelijk gaat het om een gemakkelijke vraag. Ook gemakkelijk, zij het iets minder, is vraag 2, die door 183 personen goed is beantwoord, en die dus een p-waarde heeft van 0.934 (men kan dit ook lezen als: 93% van de kandidaten heeft vraag 2 goed beantwoord). Moeilijke vragen zijn de items 5, 8, 107, en 108, waar steeds het aantal foutieve antwoorden het aantal goede overtreft.

In de vierde kolom wordt de *Rit*, voluit de *item-test (of item-totaal) correlatie*, vermeld. Deze correlatie geeft de mate weer waarin een

afzonderlijk item hetzelfde meet als de toets in zijn geheel. Het berekenen van deze correlatie tussen de score op een bepaalde vraag en de totaalscore berust op de wenselijkheid dat iedere vraag een bijdrage levert aan de totaalscore in die zin dat personen die een bepaalde vraag goed hebben, een hogere totaalscore hebben dan hun collega's die de betreffende vraag fout hebben. Men kan dat ook negatief formuleren: wanneer een vraag met name goed wordt gemaakt door overigens laag-presterende personen, dan is er met die vraag iets raars en iets onwenselijks aan de hand. De Rit is een correlatiecoëfficiënt, die kan variëren tussen -1.00 en +1.00. Waar de waarde 0 is, draagt de vraag niet bij aan het maken van onderscheid tussen hoog- en laag-presteerders. Is de waarde negatief, dan moet twijfel rijzen aan de kwaliteit van de vraag, omdat personen met een hoge totaalscore deze vraag relatief vaak fout hebben beantwoord. Bij vraag 1 is de Rit negatief, evenals bij vraag 8. Maar aan de vergelijking van deze beide vragen is goed te illustreren dat de ene Rit de andere niet is. Het is gemakkelijk in te zien dat p-waarde en Rit niet volledig onafhankelijk van elkaar variëren. Bij een p-waarde van 1.00 of 0.00 (iedereen respectievelijk niemand heeft de vraag goed beantwoord) is de Rit 0 (de vraag discrimineert niet tussen personen). Bij vraag 1 is sprake van een zeer hoge p-waarde, en de Rit van -0.028 is daar gewoon het nietszeggende artefact bij. Maar de Rit van -0.068 van vraag 8 gaat gepaard met een lage p-waarde, en het gaat dus om een moeilijke vraag die absoluut niet discrimineert tussen personen met een hoge en een lage score. Wat doet die vraag eigenlijk in die toets?, zo zou je kunnen zeggen. In ieder geval dient die vraag nader onderzocht te worden. Daarbij kunnen verrassende zaken aan het licht komen. Het kan zijn dat bij nader inzien de vraag een valstrik bevat, of eventueel gewoon onzin is. In zo'n geval dient de vraag te worden verwijderd; een tweede uitdraai bevat dan dus één (of meer) vragen minder. Niet ongewoon is overigens de bevinding dat het gaat om een fout in de sleutel: het goede antwoord heeft per ongeluk de verkeerde code gekregen. Daarmee is een triviale verklaring gevonden voor de lage p-waarde. Uiteraard dient deze vergissing te worden hersteld, en dient een tweede uitdraai te worden gemaakt.

In combinatie met de p-waarde is de Rit een belangrijk analytisch gegeven. Overigens, hoewel het bij toetsen met een groot aantal vragen nauwelijks iets uitmaakt, is een zuiverder index de *Rir*, ofwel de *item-rest correlatie*. Ieder item wordt dan niet met de totaalscore gecorreleerd, maar met de totaalscore minus de eigen bijdrage van dat item. Bij geprecodeerde toetsen is het effect hooguit kosmetisch. Maar item-analyse is ook mogelijk bij bijvoorbeeld toetsen met essayvragen, waar sprake is van minder vragen in één toets en dus de Rir de voorkeur verdient.

De vijfde kolom vermeldt het percentage personen dat de betreffende vraag heeft overgeslagen. Als regel is dat percentage 0.0. Waar dat niet het geval is, moet via de scorematrix worden gecontroleerd of de persoon in kwestie echt niets heeft ingevuld. Soms betreft het namelijk een onduidelijk ingevuld antwoord op een optisch leesbaar antwoordformulier; hetgeen hersteld dient te worden. Komt het vaak voor dat een item niet is ingevuld, dan was welllicht de toetstijd iets aan de krappe kant.

De zesde en zevende kolom, de voorlaatste en de laatste, tenslotte bieden de antwoordpercentages per antwoord-alternatief. Bij een toets met driekeuzevragen zou er een achtste kolom zijn geweest, maar hier gaat het zoals gezegd om een toets met juist-onjuist vragen. Het correcte alternatief is voorzien van een asterisk, en waar zo'n sterretje staat komt het getal overeen met de p-waarde (maal 100). Het alternatief dat niet van een sterretje is voorzien komt overeen met de *a-waarde* en geeft dus de proportie studenten weer die het verkeerde antwoord gekozen hebben. .

Tabel 3 wordt afgesloten met een reeks gegevens die na het voorgaande gedeeltelijk voor zichzelf spreken. De groep kandidaten heeft gemiddeld 77.24 punten gescoord (standaarddeviatie 5.872) op een toets die uit 100 vragen bestond. Nu vermeldt tabel 3 als nummer van de laatste vraag het getal 109, hetgeen betekent dat in het proces van item-analyse 9 vragen achteraf zijn verwijderd. Dat kan een reeks van oorzaken hebben. Bijvoorbeeld: studenten hebben tegen een of meer vragen bezwaar aangetekend, of een vraag had een negatieve item-totaal correlatie, of een vraag bleek een strikvraag te zijn, of een vraag had bij nader inzien meer dan één goed antwoord. Hoe het ook zij, het aantal vragen van de toets die wij hier bespreken was 100; daarmee is (toevalligerwijs) de gemiddelde moeilijkheidsgraad gelijk aan de gemiddelde score, met een verschuiving van de komma twee plaatsen naar links. De item-totaal correlatie bereikt bij deze toets een significantie op 5% niveau bij de waarde 0.139.

Of Kuder en Richardson twee vrouwen waren, of twee mannen, of enig andere permutatie, weet ik niet. Maar zij hebben menig formule geproduceerd, en hun *KR-20* staat voor de twintigste in de rij. Het betreft hier de betrouwbaarheid (reliability) van toetsen. Een toets is betrouwbaar wanneer de meting op niet-toevallige, consistente wijze is verkregen. Een manier om dat aan te tonen zou kunnen zijn het enige tijd later opnieuw afnemen van dezelfde toets bij dezelfde groep, of het voorleggen van een parallelle (equivalente) toets. Dat stuit uiteraard in de onderwijssituatie op praktische bezwaren. Men zou ook de toets op enige manieren in twee helften kunnen splitsen, en deze met elkaar kunnen correleren. Maar de waarde van de verkregen correlatie zal variëren met de wijze waarop de splitsing wordt uitgevoerd (de eerste versus de tweede helft, of de oneven tegen de even items). De KR-20 nu ondervangt de bezwaren van bovenstaande methoden. Het is een schatting van de homogeniteit, dat wil zeggen van de onderlinge samenhang van de items. De KR-20 wordt hoger wanneer de lengte van de toets toeneemt, en ook wanneer de groep kandidaten heterogener is. De statistische details blijven hier achterwege, maar van praktisch belang is de eis dat de waarde van de KR-20 niet beneden 0.75 mag liggen wanneer de toets het nemen van beslissingen over personen beoogt.[3] De toets van tabel 3 had dus eigenlijk niet als tentamenresultaat mogen worden gebruikt (in feite is dat wèl gebeurd, en studenten hadden daar dus tegen kunnen protesteren).

De *Signal-Noise-ratio* (de signaal/ruis-verhouding) is de verhouding tussen de ware variantie en de foutenvariantie. Omdat hij varieert met de lengte van de toets, is deze index - die ook wel met F wordt aangeduid - te gebruiken om te schatten of met de feitelijk afgenomen toets de gewenste betrouwbaarheid überhaupt te bereiken was. Ik bespaar u opnieuw de statistische details, maar de KR-20 is gelijk aan F/(F+1). Een KR-20 van .80 komt dus overeen met een F van 4.0. In tabel 3 is F 1.6427, hetgeen betekent dat de toetslengte met 4.0/1.6427 = 2.4 moet worden vermenigvuldigd voor een voldoende betrouwbaarheid. Er hadden dus 240 van dergelijke vragen gesteld moeten worden, òf men had een betere toets moeten ontwerpen.

Een bepaalde score op een toets is niet een rotsvast gegeven. Zou een ander maar equivalent tentamen zijn afgenomen, dan had dezelfde persoon waarschijnlijk een andere (hogere of lagere) score gehad. De *standaardmeetfout (smf)* nu is een schatting van het gebied waarin de 'ware score' zal liggen. Wanneer iemand 59 punten heeft verzameld op een toets met een standaardmeetfout van 3.6, dan zal zijn ware score in twee derde van de gevallen liggen tussen 55.4 en 62.6. Wijnen heeft ruim twintig jaar geleden de standaardmeetfout verwerkt in een elegante methode om de cesuur (dat is de grens tussen de voldoende en de onvoldoende prestatie) bij tentamens te bepalen.[4] Zijn centrale stelling is dat de gemiddelde groepsprestatie de beste schatting is van wat onder gegeven omstandigheden met een gegeven groep

bij een gegeven docent mogelijk blijkt. Voor individuen in een groep komt het er dus op aan om tot de groep te horen, dat wil zeggen niet af te wijken van de gemiddelde groepsprestatie. Door de grens tussen slagen en zakken te leggen bij het gemiddelde verminderd met tweemaal de standaardmeetfout, wordt de kans dat iemand ten onrechte onder die grens valt zeer klein (dat wil zeggen minder dan 5%).

De score

Wanneer het produceren van het goede antwoord 1 punt oplevert, is de score van een persoon gelijk aan het aantal goede antwoorden. Er bestaan overigens meer gecompliceerde scoringsregels (zoals bijvoorbeeld goed-minus-fout; het aantal foute antwoorden wordt in mindering gebracht op het aantal goede), maar wij beperken ons hier tot de eenvoudigste variant.

Bij geprecodeerde toetsen is de aanvankelijke score in principe een ruwe score; immers, in het resultaat zit mogelijk een zekere gisfactor verdisconteerd. Zo is bij vierkeuzevragen de a priori kans op een goed antwoord een op vier. In de gegeven uitdraai (tabel 4) valt over de scores van de personen het volgende te zeggen. Uiterst links staat de ruwe score, dat wil zeggen het resultaat in aantal uitgereikte punten. Het gaat om een toets van oorspronkelijk 109 vragen; 9 werden ex post facto verwijderd, zodat er 100 vragen overbleven, elk goed voor 1 punt. De hoogst behaalde score is 90, de laagste 56. Uiterst rechts staat het aantal personen dat een bepaalde score heeft gehaald. Een score van 90 is behaald door 1 persoon, en 2 personen presteerden 1 puntje lager. Aan de staart van het peloton bungelen 5 personen; 4 daarvan scoorden 64 punten, terwijl de hekkesluiter daar nog royaal onder blijft. De tweede kolom geeft de ruwe score weer als percentage van de maximale score. Bij 100 vragen is dat een gemakkelijke rekensom; zouden er 200 vragen zijn geweest, dan stond bij een score van 90 in de tweede kolom het getal 45.0. De derde kolom bevat het cumulatieve percentage personen dat een bepaalde score heeft behaald. Er waren 196 deelnemers aan dit tentamen; 1 kandidaat (0.5%) scoorde 56 punten, 5 (2.6%) kandidaten scoorden 64 of minder punten, en zo vervolgens.

Tabel 4. Een voorbeeld van een frequentieverdeling

Ruwe score	Procentuele score	Cumulatief percentage	Aantal personen
90	90.0	100.0	1 x
89	89.0	99.5	2 xx
88	88.0	98.5	6 xxxxxx
87	87.0	95.4	2 xx
86	86.0	94.4	3 xxx
85	85.0	92.9	3 xxx
84	84.0	91.3	8 xxxxxxxx
83	83.0	87.2	14 xxxxxxxxxxxxxx
82	82.0	80.1	9 xxxxxxxxx
81	81.0	75.5	10 xxxxxxxxxx
80	80.0	70.4	15 xxxxxxxxxxxxxxx
79	79.0	62.8	14 xxxxxxxxxxxxxx
78	78.0	55.6	15 xxxxxxxxxxxxxxx
77	77.0	48.0	10 xxxxxxxxxx
76	76.0	42.9	13 xxxxxxxxxxxxx
75	75.0	36.2	7 xxxxxxx
74	74.0	32.7	14 xxxxxxxxxxxxxx
73	73.0	25.5	10 xxxxxxxxxx
72	72.0	20.4	7 xxxxxxx
71	71.0	16.8	8 xxxxxxxx
70	70.0	12.8	4 xxxx
69	69.0	10.7	7 xxxxxxx
68	68.0	7.1	5 xxxxx
67	67.0	4.6	1 x
66	66.0	4.1	1 x
65	65.0	3.6	2 xx
64	64.0	2.6	4 xxxx
63	63.0	0.5	0
62	62.0	0.5	0
61	61.0	0.5	0
60	60.0	0.5	0
59	59.0	0.5	0
58	58.0	0.5	0
57	57.0	0.5	0
56	56.0	0.5	1 x

Een grondige bespreking van de eerder aangeduide gis-problematiek blijft hier achterwege. In zijn eenvoudigste vorm gaat het om het volgende. Stel, iemand weet met zekerheid

het antwoord op 40 van 100 tweekeuzevragen. Dat levert dus een voorlopige score op van 40 punten. De overige vragen worden gissenderwijs beantwoord, uiteraard in de helft van de gevallen met een positief resultaat. Dat levert nog eens 30 punten op, waarmee de uiteindelijke score 70 wordt. Dat resultaat is dus niet wat het lijkt, en het spreekt vanzelf dat overwegingen rond raden een rol spelen bij het bepalen van de cesuur.

Literatuur

1. Battus H. Rekenen op taal. Amsterdam: Querido, 1983.
2. Hubbard JP, Clemans WV. Multiple-choice examinations in medicine. A guide for examiner and examinee. Philadelphia: Lea & Febiger, 1961.
3. De Groot AD, van Naerssen RF. Studietoetsen construeren, afnemen, analyseren. Den Haag: Mouton, 1969.
4. Wijnen WHFW. Onder of boven de maat. Een methode voor het bepalen van de grens voldoende / onvoldoende bij studietoetsen. Academisch proefschrift. Groningen.

Aanbevolen literatuur

Dousma T, Horsten A. Tentamineren. Onderwijskundige informatie voor het Hoger Onderwijs. Utrecht: Het Spectrum, 1980.

De Groot AD, van Naerssen RF. Studietoetsen construeren, afnemen, analyseren. Den Haag: Mouton, 1969.

Opdrachten

1. U bent een vaardig medisch docent, uw inspanningen in het onderwijs mogen er zijn, u acht uw vak een van de belangrijkste zo niet hèt belangrijkste vak van de opleiding, en de resultaten van uw tentamen (100 tweekeuzevragen) komen toevalligerwijze precies overeen met die van tabel 3 en tabel 4. Verder bent u voor niets of niemand benauwd, en u hebt bekendgemaakt dat 'de norm ligt bij 61% reële kennis'. U bedoelt met dat laatste dat die personen zullen slagen die op ten minste 61% van de vragen het juiste antwoord weten te geven, plus op de resterende vragen een met de raadkans overeenkomende hoeveelheid extra punten weten te verzamelen.
   Er zijn 196 tentaminandi. Waar ligt de cesuur, en hoeveel personen zullen uw tentamen opnieuw moeten afleggen?
2. De examencommissie nodigt u uit om eens te onderzoeken wat er gebeurt als de cesuur volgens Wijnen wordt toegepast. Hoewel u van dat soort fratsen niet veel moet hebben, toont u zich bereid om dat - geheel vrijblijvend uiteraard - eens uit te rekenen. Bij welke cesuur komt u dan uit? En hoeveel personen zouden er dan méér geslaagd zijn?
3. Welke argumenten zouden door buitenstaanders (die van uw vak natuurlijk geen kaas hebben gegeten) naar voren gebracht kunnen worden om uw cesuur beschreven bij de hierboven gestelde opdracht 1 in benedenwaartse richting aan te passen?

DE AUTEUR

W. Bender is Hoofd Bureau Onderzoek van Onderwijs Geneeskunde (Boog) aan de Rijksuniversiteit Groningen.

Correspondentie-adres:
W. Bender. Bureau Onderzoek van Onderwijs Geneeskunde, Ant. Deusinglaan 1, 9713 AV Groningen.