G.M.VERWIJNEN

FOUTE ANTWOORDEN:
BENDER VERVOLGD

G.M.Verwijnen,
huisarts
Project Evaluatie
Studieresultaten
Faculteit der Geneeskunde,
Rijksuniversiteit Limburg

Correspondentieadres:
Frans v.d. Laarplein 30
6217 NH Maastricht

In het vorige nummer van dit Bulletin verscheen een intrigerend artikel van Bender getiteld: FOUTE ANTWOORDEN: ONWETENDHEID OF DEVIANT GEDRAG Bender signaleert hierin het verschijnsel van de zogenaamde "fouten-constante" bij j/o/?-toetsen en oppert een uiterst originele hypothese als mogelijke verklaring voor dit verschijnsel: de 'deviant-gedrag'-hypothese. Helaas bleek bij een experiment ter toetsing van deze hypothese geen van de op grond van de hypothese voorspelde effecten in de resultaten waarneembaar, en blijft het verschijnsel derhalve vooralsnog raadselachtig.

Het artikel prikkelde mij tot enige snuffelwerk in de databestanden van de Maastrichtse Voorgangstoets (MVT). Een toets, laat ik het voor de volledigheid nog even vermelden, die algemeen medische kennis test op het niveau van het artsexamen en die vier maal per jaar bij de volledige studentenpopulatie van de medische faculteit wordt afgenomen en, 'last but not least', volledig bestaat uit vragen van het j/o/?-type. De resultaten van dit snuffelwerk worden in deze bijdrage gepresenteerd en besproken aan de hand van een analyse van Bender's betoog.

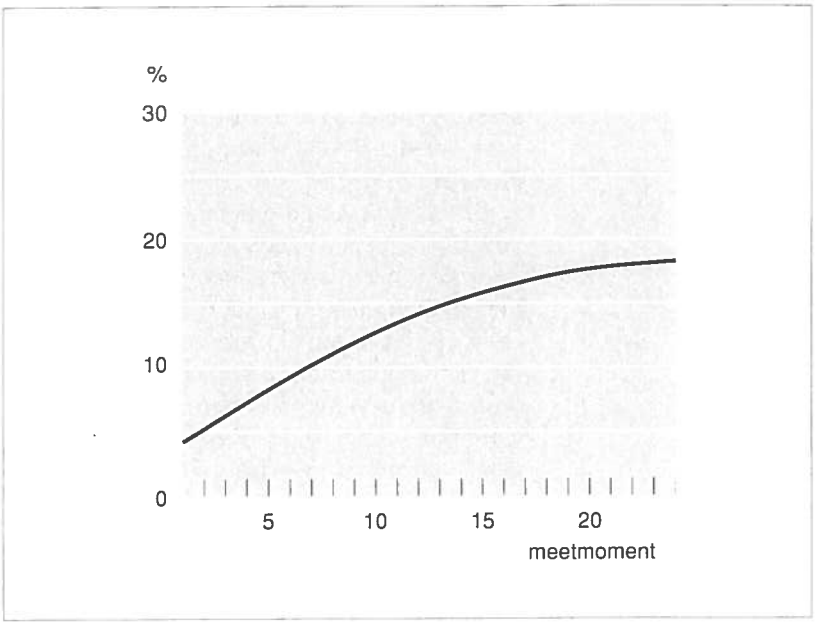
FOUTE ANTWOORDEN: HOE CONSTANT EIGENLIJK?

Om meteen maar met de deur in huis te vallen, allereerst een bespreking van het door Bender gesignaleerde fenomeen als zodanig. Is er eigenlijk wel sprake van een 'fouten-constante'? Wat leren de Maastrichtse gegevens hieromtrent?

Figuur 1 toont een samenvatting van de gemiddelde proportie foute antwoorden van een vijftal cohorten Maastrichtse studenten. Het betreft de jaargroepen die hun studie begonnen in resp. 1978, '79, '80, '81 en '82. Zoals gezegd, wordt er vier maal per jaar een voortgangstoets afgenomen, telkens in september, december, maart en mei van het betreffende academisch jaar. De gehele studie telt aldus in totaal 24 meetmomenten. De afgebeelde curve representeert de 'line of best fit' door alle datapunten bij een polynome regressie van de tweede orde. De datapunten, 5 x 24 in totaal, zijn voor de overzichtelijkheid niet afgebeeld.

Bezien we deze curve, dan springt de constantheid van de foutscores NIET onmiddellijk in het oog. Er is daarentegen sprake van een toename gedurende het studieverloop. Wel gaat de curve afvlakken in de hogere studiejaren. Mogelijk dat zich hier een plafond in het verloop aan het aftekenen is. Ogenschijnlijk wordt dit plafond omstreeks het eind van de opleiding bereikt. Zekerheid daarover ontbreekt echter, omdat we niet beschikken over vervolggegevens gedurende meerdere jaren na de opleiding. Wel hebben we gedurende circa zeven jaar elke voortgangstoets ook voorgelegd aan groepen basisartsen. Dit waren telkens zo'n 60 arts-assistenten verkerend in wisselende fasen van hun beroepsopleiding tot huisarts. Gemiddeld genomen is hun foutscore meestal net iets hoger dan die van de zesde jaars studenten op het laatste meetmoment. Alhoewel het verschil meestal statistisch significant is, kan niet gezegd worden dat het ook opzienbarend is. Gemiddeld genomen praten we over een verschil van 1%. Zesde jaars scoren aan het einde van de opleiding gemiddeld 19% foute antwoorden, basisartsen 20%.

Het zou me niet verbazen als de foutscore der basisartsen lager uitvalt dan de vermelde 20%, wanneer ook voor hen, evenals bij onze zesde jaars, 'zak/slaag' consequenties aan de toetsen verbonden zouden zijn. Ze gaan dan vermoedelijk voorzichtiger invullen en dientengevolge ook minder fouten maken.



Figuur 1.
Verloop van de foutscores

Op grond van deze gegevens meen ik dat het door Bender gepostuleerde verschijnsel van een fouten-constante op z'n minst genuanceerd moet worden, voor zover je sowieso al van een echte constante mag spreken. Uit de voortgangstoetsgegevens komt onmiskenbaar een verband tussen foutscore en opleidingsniveau naar voren. Het aantal foute antwoorden neemt toe met het stijgen der studiejaren. Wel neemt de omvang van deze toename geleidelijk af. Mogelijk dat er ooit een plafond bereikt wordt. Op dat moment is er dan vanzelfsprekend sprake van een constante. Nader onderzoek zal dit echter moeten uitwijzen. Mij ontbreken daar de gevens over en ik ben ze ook nog niet in de literatuur tegen gekomen.
Hoe het ook zij, ook mijn (aanvullende) gegevens vragen om een nadere gedachtebepaling omtrent het verschijnsel foute antwoorden. Om er simpelweg vanuit te gaan, dat er nu eenmaal ook fouten gemaakt worden bij het beantwoorden van toetsvragen, is een wat al te gemakkelijke benadering. Daarvoor is er bovendien toch bepaald teveel systematiek in de data aanwezig. In het verdere verloop van deze bijdrage zal ik proberen een eerste aanzet te doen voor deze nadere gedachtebepaling. Ik laat me daarbij nog steeds leiden door Bender's betoog.

FOUTE ANTWOORDEN: WAAROM EIGENLIJK?

Om te beginnen de vraag waarom er eigenlijk zoveel foute antwoorden gegeven worden. Naast de kwestie van de constantheid, roert Bender ook deze vraag aan en oppert een aantal verklaringen. Verklaringen die mijns inziens, in tegenstelling tot wat hij betoogt, geen van alle verworpen kunnen worden. De redeneringen op grond waarvan hij twee van de drie verklaringen buiten beschouwing meent te mogen laten, zijn mijns inziens niet voldoende steekhoudend.
Bender onderscheidt de volgende betekenissen van een fout antwoord:

1. Een fout antwoord betekent foutieve kennis.
2. Een fout antwoord betekent een onfortuinlijke gissing voortkomend uit onwetendheid.
3. Een fout antwoord betekent een onfortuinlijke gissing logisch voortvloeiend uit een inferieure vraagredactie.

De laatste twee verklaringen acht Bender niet van toepassing.

INFERIEURE VRAAGREDACTIE: TOCH EEN FOUTENBRON?

Om met de laatste verklaring te beginnen is Bender's redenering, dat het in alle gevallen zorgvuldig geredigeerde toetsen betrof met vragen waarvan de formulering ook na itemanalyse geen reden tot bezorgdheid gaf.
Zolang er geen inzicht geboden wordt in de details van de procedure, die gehanteerd wordt bij het opstellen en beoordelen van de vragen, bevredigt een dergelijke redenering mij allerminst.
Dit verklaart zich door de ervaringen hieromtrent, die wij in de loop der jaren in Maastricht hebben opgedaan. Zonder overdrijven meen ik te mogen stellen, dat de Maastrichtse gang van zaken in deze absoluut onovertroffen is. Ik ben geneigd te zeggen, dat het niet zorgvuldiger kan dan zoals in Maastricht te doen gebruikelijk is. Alhoewel ik mezelf zojuist verplicht heb om dergelijke uitspraken te doen vergezellen van een nadere toelichting, verwijs ik, omwille van het verloop van mijn

betoog, op deze plaats naar beschikbare literatuur.[1,2]
Wat leert die Maastrichtse ervaring dan wel? Heel algemeen geformuleerd komt het er op neer dat het systeem nimmer onfeilbaar is. Hoe zorgvuldig je procedure ook lijkt, je kunt er gevoegelijk van uitgaan dat er altijd blinde vlekken blijven. Altijd weer slippen er vragen door de mazen van het net, die de toets der kritiek achteraf bezien niet kunnen doorstaan. Dit is op zich natuurlijk niet zo verwonderlijk, mensen maken nu eenmaal vergissingen. Belangrijk is echter dat het niet om zomaar een enkel vraagje gaat. Ik spreek hier van een getal van gemiddeld zo'n 7% van het oorspronkelijk aantal toetsvragen. Dat is naar mijn smaak niet mis. Nog belangrijker is echter, dat deze vragen slechts sporadisch aan het licht komen middels bestudering van de gebruikelijk item-analyses. Het merendeel komt onder onze aandacht door commentaar van de studenten. Wij maken daar systematisch gebruik van. Het is een wezenlijk onderdeel van de bewakingsprocedure, naar onze mening volstrekt onmisbaar! De studenten worden nadrukkelijk aangespoord tot het leveren van commentaar. Slechts zelden is een dergelijke activiteit in de bewakingsprocedure ingebouwd. Ik neem aan dat dit ook bij de toetsen van Bender niet het geval is. Hij mag er met andere woorden niet zonder meer van uit gaan, dat vragen van inferieure kwaliteit in zijn toetsen ontbreken, louter en alleen op grond van de bewering, dat ze zorgvuldig geredigeerd waren en ook de statistische itemgegevens achteraf geen reden tot bezorgdheid gaven.
Ter illustratie geef ik hier een typisch voorbeeld van zo'n vraag uit het Maastrichtse voortgangstoetsmateriaal:

Voorbeeld 1:
Erysipelas wordt teweeggebracht door een endotoxine van (hemolytische) streptococcen.
ONJUIST

Ogenschijnlijk een eenduidige en probleemloze vraag. Toch zit er een addertje onder het gras. Het draait in deze vraag namelijk om het woordje 'endotoxine', maar dat wordt niet als zodanig benadrukt. De vraag is met andere woorden onvoldoende duidelijk toegespitst op de essentie. Dit klemt temeer, daar erysipelas - een vrij veelvuldig voorkomende, bepaald geen onschuldige huidaandoening, die goed door de huisarts behandeld kan worden - wel veroorzaakt wordt door (hemolytische) streptococcen. Daar speelt een endotoxine echter geen rol bij. Door de vraag hier onvoldoende nadrukkelijk op toe te spitsen, bestaat er een gerede kans dat er 'in de hitte van de strijd' overheen gelezen wordt, zeker wanneer men bekend is met de verwekker van erysipelas.
Je zou met andere woorden kunnen zeggen, dat de a priori kans om deze vraag verkeerd te beantwoorden vanwege de formulering sowieso hoger zal zijn dan de kans om hem goed te beantwoorden. Dit komt naar mijn smaak ook tot uitdrukking in de resultaten.
Tabel 1 bevat deze resultaten in termen van percentages studenten die deze vraag respectievelijk goed (G), fout (F) en met een vraagteken (?) beantwoordden. Voor elke jaargroep (JG) zijn de gegevens apart vermeld. De betreffende vraag is opgenomen geweest in een voortgangstoets, die ook bij een aselecte steekproef van steeds ca. 30 studenten uit elk jaar van de medische faculteit Nijmegen werd afgenomen, zodat ook die resultaten hier gepresenteerd kunnen worden.

TABEL 1

	MAASTRICHT			NIJMEGEN		
JG	G	F	?	G	F	?
1	1	2	97	0	3	97
2	1	4	95	0	0	100
3	6	17	77	8	39	53
4	6	36	58	20	50	30
5	15	66	19	13	50	37
6	22	69	9	15	65	20

Deze vraag werd destijds uiteraard uit de toets verwijderd, alvorens de definitieve scores berekend werden. Dit gebeurt in Maastricht met alle vragen, die achteraf bezien toch problematisch blijken te zijn. Hier zou dus met recht gesteld kunnen worden dat Bender's derde verklaring niet van toepassing is. Ik koester echter niet de illusie, dat er dientengevolge geen probleemvragen meer in het toetsmateriaal zijn achtergebleven. Sedert enkele jaren brengen wij eerder gebruikte vragen opnieuw in circulatie, waarbij elke vraag opnieuw 'sans rancune' de volledige beoorde-

lingsprocedure volgt. In dit materiaal komen we tot onze ontsteltenis regelmatig vragen tegen, die de toets der kritiek op generlei wijze doorstaan. Dit maant vanzelfsprekend tot voorzichtigheid ten aanzien van uitspraken omtrent de 'zuiverheid' van je toetsmateriaal. Al met al is mijn conclusie, dat op z'n minst een deel van de foutscore te verklaren is via Bender's derde hypothese, inferieure vraagredactie. In elk geval zijn er onvoldoende aanwijzingen om de hypothese te verwerpen.

EEN GOKJE WAGEN: UITGESLOTEN?

Ook de gedachte, dat de foutscore verklaard moet worden via verkeerd uitgevallen gissingen uit onwetendheid, wordt door Bender verworpen, omdat ruimschoots gebruik gemaakt wordt van het vraagteken alternatief. Om aan het ruime vraagteken gebruik de conclusie te verbinden, dat gissen inderdaad voorkomen werd, gaat me te ver, zeker wanneer we het hebben over zgn. niet-meetellende toetsen. Het vraagtekengebruik is hooguit een indicatie voor de mate van absolute onwetendheid van de kandidaten. Gokgedrag wordt met de gebruikelijke scoringsmethode (goed = +1 punt, fout = -1 punt en vraagteken telt als nul) NIET uitgebannen. De literatuur laat daarover geen twijfel bestaan. Ik hoef slechts te verwijzen naar een artikeltje dat recent verscheen in Medical Education.[3] Maar die literatuur heb ik hier niet nodig. Een ieder die ooit wel eens een objectieve studietoets maakte (wie heeft dat tegenwoordig niet gedaan?), weet uit ervaring, dat het aantal vragen waarop je het antwoord onmiddellijk en zonder enige twijfel precies weet, slechts gering in getal is. Veelal is er echter sprake van een zekere mate van onzekerheid. Soms veel soms weinig, maar lang niet altijd onaanzienlijk, waag ik te veronderstellen. Zelden is kennis op dezelfde nauwkeurige en exact omschreven wijze in het geheugen vastgelegd, als de aard van de meeste vragen doet veronderstellen. Regelmatig zullen we proberen een vraag toch te beantwoorden, ook al hebben we de exacte gegevens niet paraat. We weten niet precies (meer) hoe het (ook weer) zit, maar op grond van allerlei gedachten en veronderstellingen menen we toch een weloverwogen gokje te kunnen wagen. Een typisch voorbeeld van dit mechanisme illustreert de volgende vraag.

Voorbeeld 2:

Bij Hemophilie-A is de bloedingstijd in de meerderheid der gevallen verlengd.
ONJUIST

De redactie is onberispelijk en ook inhoudelijk is er geen enkel probleem. Je moet echter precies weten wat er aan de hand is met dit ziektebeeld en bovendien nauwkeurig beseffen wat het begrip bloedingstijd inhoudt, om deze vraag correct te kunnen beantwoorden. Op grond van algemeen heersende kennis van dit ziektebeeld zal echter de neiging om de vraag te beantwoorden groot zijn, ook al besef je dat je het niet precies (meer) weet. Het gokje is dan gauw gemaakt, de fout ook. Een verkeerd uitgevallen gissing dus.
De resultaten in tabel 2 spreken voor zich.

TABEL 2:

	MAASTRICHT			NIJMEGEN		
JG	G	F	?	G	F	?
1	1	32	67	0	16	84
2	2	40	58	5	18	77
3	6	44	50	0	47	53
4	40	40	20	27	53	20
5	55	23	22	4	79	17
6	59	22	19	23	60	17

Kortom, ik meen dat Bender's tweede verklaring voor foute antwoorden niet terzijde geschoven mag worden. Gokgedrag is niet uitgesloten. We praten echter niet over domweg willekeurig gokken. In zoverre ben ik het met Bender eens. Gissingen uit absolute onwetendheid zullen zich vermoedelijk niet vaak voordoen en dus ook geen noemenswaardige bijdrage aan de foutscore leveren. Strikt genomen wordt Bender's verklaring dan ook terecht verworpen. Ik meen echter dat gissen ruimer opgevat moet worden dan Bender suggereert. Bij veel vragen zal er sprake zijn van een zogenaamde 'educated guess', die van geval tot geval meer of minder 'educated' zal zijn. Het lijkt me niet reëel om te veronderstellen, dat in al deze gevallen de gok in de goede richting uitvalt. Ergo, een deel van de foute antwoorden vindt z'n oorsprong in verkeerd uitgevallen gissingen. Over de omvang hiervan kan slechts gespeculeerd worden. Nader onderzoek zal daaromtrent inzicht moeten verschaffen. Vooralsnog mag men er

echter niet van uitgaan dat dit een te verwaarlozen aantal is.

FOUTIEVE KENNIS OF DEVIANT GEDRAG: EEN CONSTANTE?

Over Bender's eerste verklaring, fout antwoord is foute kennis, kan ik kort zijn. Het lijkt me volstrekt logisch om te veronderstellen, dat in het proces van kennisverwerving en begripsvorming, misvattingen kunnen optreden. Misvattingen, die dienovereenkomstig aanleiding geven tot een foutief antwoord. Een deel van de foute antwoorden zal dan ook zeker langs deze weg verklaard moeten worden. Bender oppert in dit verband zijn deviant-gedrag hypothese om te verklaren waarom het volume foutieve kennis niet varieert met de gemiddelde leeftijd of, zoals ik dit interpreteer, met het kennisniveau van de participerende groepen. Aannemende dat er inderdaad sprake is van een constante, zoals Bender veronderstelt, dan ontgaat mij echter ten ene male de ratio van de redenering dat gebrek aan consensus leidt tot een min of meer gelijkblijvende hoeveelheid foute antwoorden. Ik ben eerder geneigd om te veronderstellen, dat naarmate men meer weet er ook meer verschil van mening zal zijn, en er dus ook meer aldus te verklaren foute antwoorden gegeven zullen worden. Een verklaring voor constantheid kan ik er in elk geval niet in zien. Wederom een voorbeeld ter illustratie:

Voorbeeld 3:

Chronisch recidiverende diarrhee bij een NIET ziek kind kan het gevolg zijn van een 'irritable colon'.
JUIST

De vraag is geen fraai voorbeeld van redactionele kwaliteit. Toch illustreert hij op saillante wijze wat ik bedoel. In Maastricht hanteren we namelijk de stelling, dat zoveel mogelijk vermeden moet worden om het woordje 'kan' te gebruiken. Vrijwel alles kan in de geneeskunde en een ONJUIST sleutel bij dergelijke vragen houdt zelden stand. Zo'n vraag 'triggert' met andere woorden de voorkeur voor het antwoord JUIST bij de respondenten. De kans op fouten is dus gering. Zo ook bij deze vraag, getuige de gegevens van met name de jongere jaars studenten (zie tabel 3). Ofschoon een 'irritable colon' inderdaad beschreven wordt als een mogelijk oorzaak van chronisch recidiverende diarrhee, is dit in de praktijk slechts zelden het geval. In verreweg de meeste gevallen van dit in de eerste lijn veelvuldig voorkomende ziektebeeld, ligt er een andere oorzaak aan ten grondslag. Dit begint pas echt tot je door te dringen als je er praktisch mee te maken krijgt, zodanig zelfs dat je geneigd bent te veronderstellen, dat een 'irritable colon' geen oorzaak kan vormen. Je verklaart je dan oneens met de stelling en vertoont in Bender's zin deviant gedrag. Dit verklaart mijns inziens de toegenomen percentages foute antwoorden in de hogere studiejaren. Deze vraag werd destijds ook voorgelegd aan huisartsen in opleiding (HAIO's), zodat ik die resultaten ook in de tabel heb kunnen opnemen. Dan begint men pas echt in de gaten te krijgen hoe het in de praktijk is. Geen wonder dat een duidelijke meerderheid nu deviant gedrag gaat vertonen.

TABEL 3:

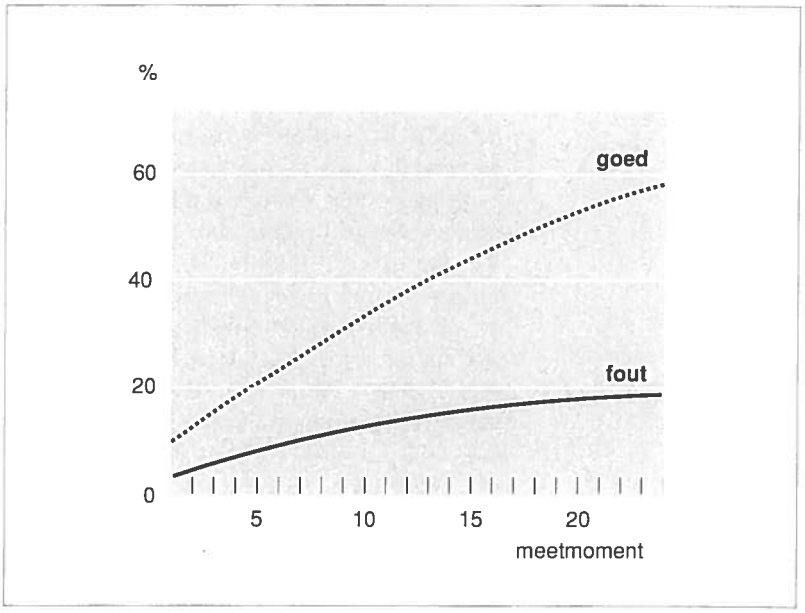
	MAASTRICHT			NIJMEGEN		
JG	G	F	?	G	F	?
1	41	1	58	13	6	81
2	44	2	54	27	0	73
3	52	8	40	53	0	47
4	60	14	26	63	10	27
5	49	34	17	50	13	37
6	57	22	21	62	25	13
HAIO's	33	60	7			

Recapitulerend dus: deviant gedrag als verklaring voor foute antwoorden, accoord, maar niet als verklaring voor een constante.

FOUTE ANTWOORDEN: HOE ZIT HET DAN?

In het voorgaande heb ik geprobeerd te beargumenteren, dat elk van Bender's verklaringen een zekere geldigheidswaarde heeft. Bovendien heb ik gegevens aangedragen, waaruit blijkt dat het percentage foute antwoorden GEEN constante is. Het varieert op z'n minst met het opleidingsniveau van de respondenten. Maar naarmate het opleidingsniveau stijgt, worden de variaties kleiner. Er is

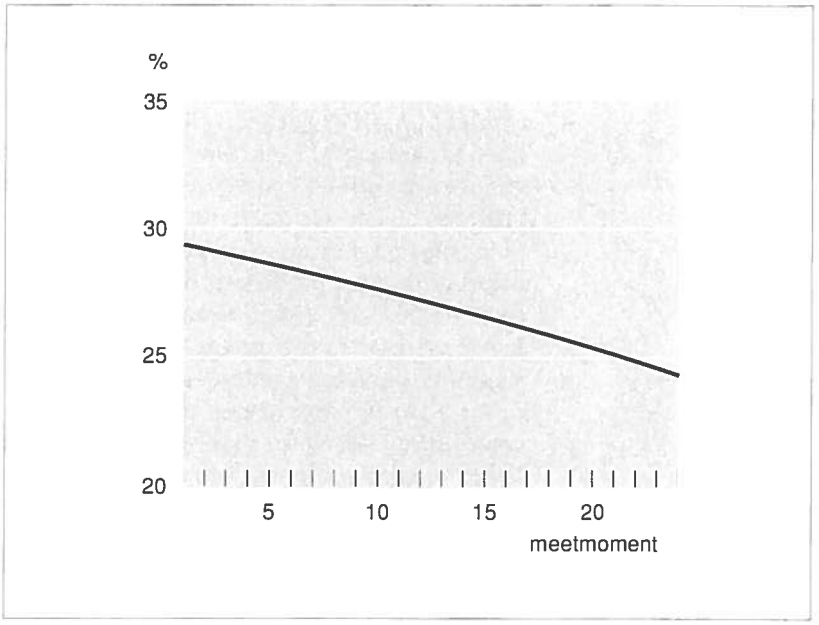
een indicatie dat de foutscore een zekere maximumwaarde niet zal overschrijden. Deze maximumwaarde ligt gemiddeld genomen om en nabij de 20% van de totaal te behalen score. Persoonlijk ben ik geneigd om te veronderstellen, dat het merendeel van de foutscore berust op mechanismen die verband houden met Bender's tweede verklaring. Fouten zijn merendeels verkeerd uitgevallen 'educated guesses'. De oorzaak van het verkeerd uitvallen van deze gissingen berust op een interactie tussen de kwaliteit van de vraagformulering, het kennisniveau van de respondenten en, 'last but not least', de onwillekeurige voorkeursneiging voor een bepaald antwoord bij gissen. Op dit laatste verschijnsel kom ik later nog terug. De interactie tussen deze variabelen is complex en wisselend en laat zich derhalve niet gemakkelijk bestuderen. Het verbaast me niets dat Bender's experiment geen eenduidige gegevens opleverde. Niet omdat de deviant-gedrag hypothese ondeugdelijk is. In tegendeel, het lijkt me een uiterst plausibele verklaring voor het ontstaan van foute antwoorden. Dat heb ik trachten te illustreren aan de hand van mijn laatste voorbeeld.
Het is echter niet de enige verklaring en daar heeft Bender onvoldoende rekening mee gehouden in zijn experiment. Vandaar vermoedelijk de weinig eenduidige resultaten.
Om de invloed van een der opererende variabelen vast te stellen, zou men de overige liefst constant houden. Dat is in dit verband meestal praktisch onmogelijk. Vooralsnog moeten we daarom volstaan met zorgvuldige exploratie en interpretatie van de beschikbare gegevens. Mogelijk dat daar uiteindelijk experimenteel toetsbare hypothesen uit af te leiden zijn.



Figuur 2.
Verloop van goed- en foutscores.

SNUFFELEN VOORTGEZET

In figuur 2 is nogmaals het verloop van de foutscores afgebeeld, doch ditmaal in samenhang met het verloop der goedscores. Wederom de 'line of best fit' door de gemiddelde proportionele goedscores van de eerder genoemde vijf cohorten. Er is bepaald enige overeenkomst in het scoreverloop. De curve van de goedscores loopt echter veel steiler omhoog en er is niet zo duidelijk sprake van een afvlakking in de hogere studiejaren. In feite neemt dus de proportie foute antwoorden van het totaal aantal beantwoorde vragen (F/G+F) in de loop der studiejaren af. Dit verloop is afgebeeld in figuur 3. Vatten we de goedscores op als een maat voor kennis, dan kunnen we de gegevens uit figuur 2 als volgt interpreteren: In het begin van de kennisontwikkeling neemt het percentage foute antwoorden toe met de toename van de kennis, maar naarmate de kennisontwikkeling vordert, vermindert of verdwijnt dit verband. Betrek ik ook figuur 3 hierbij, dan rijst de vraag of je dit ook mag opvatten als: 'In het begin geldt: hoe meer goed, hoe meer fout, maar naarmate het kennisniveau stijgt, raakt fout steeds onafhankelijker van goed.' We praten hier echter over gemiddelde resultaten. Om deze vraag te beantwoorden moeten



Figuur 3.
Verloop proportie fout der beantwoorde vragen.

individuele resultaten bestudeerd worden. Het gaat dan om de correlatie tussen de goed- en de foutscores. Gesteld dat het zojuist geopperde verband opgaat, dan zou moeten gelden, dat de correlatie tussen de goed- en de foutscores met het vorderen der studiejaren afneemt van relatief hoog bij jongere jaars studenten tot relatief laag of nihil bij oudere jaars. Tabel 4 bevat deze correlaties bij een tweetal voortgangstoetsen. De voortgangstoets van mei 1987 en mei 1988. Ik koos voor de toets van mei '88, omdat we deze toets nog eens eenmalig aan een grote groep (n=179) recent afgestudeerde basisartsen hebben voorgelegd, zodat ik hier ook de resultaten van deze groep kan presenteren. Sinds 1984 hebben we namelijk geen voortgangstoetsen meer bij basisartsen afgenomen, maar bij deze gelegenheid is dit nog eens herhaald om na te gaan of de resultaten van destijds nog steeds gelden. Ter vergelijking is de overeenkomstige voortgangstoets van het jaar daarvoor genomen.

TABEL 4:

JG	1987	1988
1	0.87	0.77
2	0.69	0.56
3	0.27	0.34
4	0.09	0.09
5	0.16	0.01
6	0.16	0.13
BA		0.08

De resultaten zijn overeenkomstig de verwachting. Ergo, bij jongere jaars is er een duidelijk positief verband tussen het aantal goed en het aantal fout beantwoorde vragen en deze positieve correlatie daalt en gaat verloren bij het vorderen der jaren.
Hoe is dit te verklaren? Ik kom nu terug op mijn eerder geopperde veronderstelling omtrent de oorzaak van foutscores. Foutscores hebben te maken met gokgedrag. Foute antwoorden berusten merendeels op verkeerd uitgevallen weloverwogen gissingen ('educated guesses'). Deze gissingen vallen verkeerd uit als gevolg van een complexe interactie tussen vraagredactie, kennisniveau en onwillekeurige antwoordneiging (de engelse vakterm hiervoor is: 'response set') van de respondent. Om met dit laatste te beginnen het volgende. De 'founding father' van de 'true/false'-vraag Ebel, schreef reeds in zijn handboek: "in the absence of firm knowledge a student seems more likely to accept than to question a declarative statement whose truth or falsity he must judge".[4] Dit zou betekenen, dat, indien een aanzienlijk deel der vragen beantwoord wordt 'in the absence of firm knowledge', er gemiddeld genomen vaker voor het antwoord JUIST gekozen wordt dan voor het antwoord ONJUIST. Bij JUIST gesleutelde vragen (vragen met JUIST als het correcte antwoord) is dit een GOED antwoord, bij ONJUIST gesleutelde vragen echter FOUT. Met andere woorden bij ONJUIST gesleutelde vragen zal de gissing vaker verkeerd uitvallen dan bij JUIST gesleutelde vragen. De gemiddelde foutscore van de ONJUIST gesleutelde vragen zal dientengevolge hoger zijn dan de gemiddelde foutscore der JUIST gesleutelde vragen. Dit is inderdaad het geval getuige de resultaten weergegeven in tabel 5. Het betreft in alle gevallen statistisch significante verschillen ($P \leq 0.01$).

TABEL 5:
Gemiddelde foutscore van 1e tot en met 6e jaars voor JUIST (j) en ONJUIST (o) vragen bij twee Maastrichtse Voortgangstoetsen (M1 en M2). De tweede toets werd ook in Nijmegen afgenomen (N)

JG	sleutel	M1	M2	N
1	j	5	5	7
	o	13	13	15
2	j	8	7	8
	o	15	15	14
3	j	9	10	14
	o	19	18	21
4	j	12	12	16
	o	24	21	25
5	j	12	15	20
	o	27	24	28
6	j	13	15	17
	o	24	23	23

De conclusie is duidelijk. Het door Ebel beschreven fenomeen speelt inderdaad een statistisch significante rol bij de totstandkoming van de toetsresultaten. Zoals ik al eerder zei, mag je dus ook aannemen dat er inderdaad op vrij grote schaal 'gegokt' wordt. Vanwege de

overwegend in bevestigende zin werkende 'response set' zullen verkeerd uitvallende gissingen voornamelijk bij de ONJUIST gesleutelde vragen voorkomen. Ofschoon dit gemiddeld genomen het geval moge zijn, betekent dit geenszins dat het voor elke respondent bij elke vraag op dezelfde wijze werkt. Er zullen zeker studenten zijn, bij wie de 'response set' precies andersom is afgesteld. Ze zijn waarschijnlijk echter in de minderheid.
Bovendien moet ook nog rekening gehouden worden met effecten van de vraagredactie en het kennisniveau van de studenten. Ik wil dit illustreren aan de hand van twee laatste voorbeelden. Om met de vraagredactie te beginnen.

voorbeeld 4:
Een vrouw met zwangerschapsdiabetes wordt op dezelfde wijze gecontroleerd en behandeld als een zwangere, die vóór haar graviditeit al manifeste diabetes mellitus had.
JUIST

Op grond van bovenbeschreven fenomeen zou je mogen verwachten, dat de vraag vaker goed dan fout beantwoord wordt. Het is immers een JUIST gesleutelde vraag. Het tegendeel is echter het geval. Zie de resultaten in tabel 6. Dit is mijns inziens voornamelijk het gevolg van de ongelukkige redactie, waardoor je, ook zonder enige kennis van zaken, welhaast onherroepelijk naar het antwoord ONJUIST getrokken wordt.

TABEL 6:

	MAASTRICHT			NIJMEGEN		
JG	G	F	?	G	F	?
1	9	15	76	13	39	48
2	24	38	38	23	41	36
3	29	43	28	22	44	34
4	50	36	14	47	50	3
5	32	55	13	46	50	4
6	46	46	8	50	42	8

Het volgende en laatste voorbeeld grijpt terug op de vraag die gebruikt is in voorbeeld 2. Hetzelfde onderwerp is eens op iets andere wijze in een andere toets aan de orde geweest. Het betrof een september-toets, dus aan het begin van het studiejaar. De studenten zijn dan inmiddels drie tot vier weken bezig, sinds de vakantieperiode. De vraag luidt aldus:

voorbeeld 5:
Bij Haemophilie A wordt bij laboratorium onderzoek in het merendeel der gevallen een verlengde bloedingstijd gevonden.
ONJUIST

Om twee redenen neigt de vraag tot een fout antwoord (dat is hier dus JUIST). De eerste reden werd reeds besproken bij voorbeeld 2. De tweede reden is de verkeerd uitvallende 'response set' bij een ONJUIST gesleutelde vraag. Geen wonder dat er reeds bij jongere jaars opvallend veel foute antwoorden gegeven worden als ze de vraag beantwoorden. Zie tabel 7.

TABEL 7:

JG	G	F	?
1	3	29	68
2	2	60	38
3	3	57	40
4	72	26	2
5	43	48	9
6	55	34	11

Opvallend is echter dat de vierde jaars plotseling zo vaak een goed antwoord geven. Welnu, toevallig is het zo dat het eerste blok in het vierde jaar in de tijd dat deze toets werd afgenomen, handelde over het thema bloedverlies. Deze studenten zitten goed in de leerstof en beantwoorden de vraag daarom merendeels goed. Kennelijk wordt de onwillekeurige neiging om de vraag verkeerd te beantwoorden, deels als gevolg van de 'response set' en deels vanwege het effect van de inherente logica bij partiële kennis, hier overheerst of verdrongen door goede kennis van zaken. Erg krachtig en duurzaam is dit mechanisme echter niet, getuige de resultaten van de volgende jaargangen, die destijds ook dit blok in hun vierde jaar volgden. Bezien we nog eens de resultaten bij voorbeeld 2 (tabel 2), dan is er zelfs al vrij snel sprake van verval. De betreffende vraag was destijds opgenomen in een maart-toets, dat wil zeggen ongeveer een half jaar nadat het thema bloedverlies in het onderwijs voor de vierde jaars centraal stond.
Wat ik met dit laatste voorbeeld heb willen illustreren is, dat er globaal genomen kennelijk sprake is van twee elkaar min of meer

beconcurrerende mechanismen, die bepalend zijn voor de uitslag van het keuzeproces, wanneer een vraag beantwoord gaat worden. Toetsdeskundigen spreken hier van latente trekken, min of meer verborgen eigenschappen van de respondenten, die in de testresultaten tot uitdrukking komen. De ene latente trek is in dit geval *feitenkennis*, de trek die men ook bedoeld te meten, de ander is de *onwillekeurige antwoordneiging* van de kandidaat, waar de toetsconstructeur natuurlijk niet op uit is, maar die echter wel in het geding komt. Dit zijn geen statische kenmerken, maar voortdurende veranderende fenomenen. Feitenkennis is onderhevig aan groei en verval. Voor de onwillekeurige antwoordneiging ligt dit iets anders. Min of meer vaste persoonlijkheidskenmerken zullen zeker van invloed zijn. Je bent een durfal of niet. De mate waarin deze kenmerken tot uitdrukking komen zal echter van vraag tot vraag verschillen. Daar zijn allerlei factoren van invloed op, zoals bijvoorbeeld vraagredactionele factoren. Maar ook, en daar gaat het me hier om, de veranderingen in de feitenkennis. Er is sprake van een dynamische evenwichtsreactie tussen deze beide trekken. Afhankelijk van de hoeveelheid feitenkennis is deze reactie meer of minder naar rechts of links verschoven. Hoe meer feitenkennis en hoe beter deze is omschreven en vastgelegd in het geheugen, hoe minder de onwillekeurige antwoordneiging tot uitdrukking kan komen, hoe meer dus een correct antwoord een weerspiegeling is van feitenkennis en hoe minder kans op een verkeerd uitvallende gissing.
Dit nu opper ik als verklaring voor de dalende correlaties tussen de goed- en de foutscores met het vorderen der jaren. Door gebrek aan welomschreven kennis bij jongere jaars studenten wordt de onwillekeurige antwoordneiging nog weinig onderdrukt en komt in min of meer gelijke mate tot uitdrukking in zowel de goed- als de foutscores. Ergo hoge correlaties. Naarmate het kennisniveau echter stijgt wordt deze trek steeds beter beheerst. De onwillekeurige antwoordneiging komt relatief steeds minder tot uitdrukking in de goedscores, terwijl deze nog steeds voor een aanzienlijk deel der foutscores verantwoordelijk is, waarbij, ik herhaal het nog maar eens, met name de invloed van vraagredactie maar ook van foutieve kennis casu quo 'deviant-gedrag' zich doet gelden. Ergo, de correlaties tussen de goed- en de foutscores dalen tot nihil. Er ontstaat uiteindelijk een soort evenwichtssituatie, waarbij het aantal foute antwoorden, maar ook het aantal goede antwoorden binnen zekere grenzen constant blijft.

SLOTBESCHOUWING: VERWIJNEN VERVOLGD!

Ik heb in deze bijdrage getracht aan te tonen, dat Bender's beschouwing omtrent het fenomeen foute antwoorden op diverse plaatsen mank gaat. Er is geen sprake van een constante. Niet alleen goedscores maar ook foutscores nemen toe met het toenemen van kennis, alhoewel dit mogelijk aan een zeker maximum gebonden is. Verklaringen omtrent de betekenis van foute antwoorden mogen geen van alle verworpen worden, ook niet Bender's 'deviant-gedrag' hypothese, ondanks de tegenvallende resultaten van zijn experiment. Aan de hand van enige aanvullende gegevens uit voortgangstoetsmateriaal opper ik een alternatieve verklaring omtrent de grondslag van foute antwoorden. Gokgedrag vormt de basis voor een ingewikkelde interactie tussen feitenkennis, vraagredactie en onwillekeurige antwoordneiging, die uiteindelijk bepalend is voor de scores. Niet alleen foutscores, maar ook goedscores!
Een en ander roept eens te meer twijfels op omtrent de geldigheid van de interpretaties en consequenties die we gewoon zijn aan toetsresultaten te verbinden. Ik heb daar eerder al eens op gewezen tijdens een NVMO-studiedag.[5] Toen ging het om het verschil tussen JUIST- en ONJUIST gesleutelde vragen en de merkwaardig lage correlaties tussen deze onderdelen van één en dezelfde toets. Ook toen vraagtekens bij wat we eigenlijk aan het doen zijn met onze toetserij. De psychometrie toont ons echter voortdurend, dat onze toetsen uitstekend voldoen. Ze beantwoorden aan het doel waarvoor ze ontworpen zijn, de resultaten zijn nauwkeurig en reproduceerbaar.
Maar toch! Telkens weer doen zich onverwachte verschijnselen voor, die niet zonder meer te verklaren zijn, die manen tot voorzichtigheid en die in de laatste maar zeker niet de minste plaats vragen om nader onderzoek. Ook al ben ik het dan niet eens met Bender's beschouwing, hij brengt de gemoederen tenminste weer in beweging, in elk geval de mijne. Dat moge blijken uit deze bijdrage. Ik

kwam hier met aanvullende gegevens en deed een poging om deze te duiden. Het antwoord is er echter nog niet. Ook Verwijnen zal nog vervolgd moeten worden.

LITERATUUR

1. Hessen Van PAW, Verwijnen GM. Necessity of test review committee in test construction. Paper gepresenteerd op het International Symposium on evaluation in medical education. 1987 Beer Sheva, Israël.

2. Verwijnen GM, Fröberg-Gresnich C. Jaarverslag voortgangstoetsbeoordelingscommissie. PES-bulletin 1986; 117. Rijksuniversiteit Limburg, Maastricht.

3. Fleming PR. The profitability of 'guessing' in multiple choice question papers. Medical education 1988; 22: 509-513.

4. Ebel RL. Essentials of educational measurement. Englewood Cliffs, Prentice-Hall, 1972.

5. Verwijnen GM, Imbos Tjl. Ja/nee-vragen: toch nog een paar vraagtekens. Paper gepresenteerd op de jaarlijkse NVMO-studiedag. 1987, Amsterdam.

STAGEPERIKELEN

Sjouk en ik liepen stage in dezelfde plaats, in het tweede jaar van de opleiding maatschappelijke gezondheidszorg. We werkten in verschillende wijken en we ontdekten al snel dat onze stagebegeleidsters dikke vriendinnen waren. Niet alleen droegen zij het haar op dezelfde lengte, met links een scheiding en rechts een schuifspeldje, zij reden ook beiden hetzelfde type auto, uitgevoerd in de kleur bordeauxrood. Naderhand, toen we bij hen thuis kwamen voor een werkbespreking, bleken zij in een flat te wonen, precies tegenover elkaar. De een woonde op nummer vierendertig, de andere op nummer tweeëndertig. Maar of je nou bij de een binnenstapte of bij de ander, dat maakte geen verschil, want de inrichting van beide flats was vrijwel identiek. Vanwege hun tijdloze kleding, waarbij de enige variant was dat de ene een beige regenjas droeg, terwijl die van de ander uitgevoerd was in een onbestemde kleur grijs, was hun leeftijd moeilijk te schatten. Zij konden zowel dertig jaar zijn als veertig of misschien lag het er ergens tussenin. We zijn er nooit achter gekomen. Overigens viel er niets op hen aan te merken. Het waren wijkzusters van het oude stempel. Rust, Reinheid en Regelmaat wat betreft de verpleging. Ging het om voorlichting voor een verantwoorde leefwijze, dan hanteerden zij de Schijf van Vijf. Geen drank en sigaretten uiteraard.
Nu hadden Sjouk en ik beiden een gedegen opleiding achter de rug. Wij hadden in ziekenhuizen gewerkt, waar een gesprek met een patiënt alleen werd toegestaan als zulks gecombineerd werd met nuttige werkzaamheden, zoals het soppen van bedden, het dweilen van zalen of het opruimen van de linnenkast. We schikten ons naar de werkmethoden van onze stagebegeleidsters en bespraken 's avonds thuis de creatieve aspecten van ons werk en de boeiende menselijke kanten ervan. Zo kwamen wij toch aan onze trekken.
Als ik een middag op het zuigelingenbureau meedraaide, wees mijn stagebegeleidster mij gretig op moeders die een gestreken luier uit hun tas haalden, met een blik van: "Zie, dat zijn nu de echte moeders, het goede soort, die strijken hun luiers nog!" Behalve dat zij werk van ons hadden doordat zij beoordelingen over ons moesten maken, kregen zij na verloop van tijd ook plezier van ons, want we namen toch ongeveer de helft van hun werk over. In het kader van "Oefening baart kunst" kregen we zeker niet de makkelijkste patiënten toegeschoven. Tenslotte gingen zij (samen uiteraard) ook nog een week op vakantie, zodat wij de zaak alleen moesten runnen.
Maar ziet, aan het einde van onze stageperiode werd ons gezwoeg dan toch beloond. Na een stagebeoordeling, diezoals te verwachten, ook vrij kleurloos was- keken de dames ons geheimzinnig aan en werd ons voor de week daarna een etentje aangeboden in een visrestaurant. Nu houd ik niet van vis, maar het was een dermate positief gebaar, dat het suggereren van alternatieven allerminst op zijn plaats was.
Sjouk en ik hadden van tevoren reeds overlegd dat, voor zover wij de dames kenden, het niet raadzaam was iets uit te zoeken in de categorie dure vissoorten. Toen wij dan ook in een van de bordeauxrode auto's naar het visrestaurant waren gereden en ons aldaar geïnstalleerd hadden, kozen Sjoukje en ik prompt het op een na goedkoopste gerecht, een ordinaire kabeljauw met botersaus, gecombineerd met worteltjes.
Toen keek mijn stagebegeleidster ons beiden aan. "Wat willen jullie", vroeg ze en pauzeerde even om goed tot ons door te laten dringen wat er vervolgens komen ging: "Willen jullie nu iets bij het eten drinken óf hebben jullie liever straks een kopje koffie toe?" Mijn mond zakte open en ik moet geleken hebben op de vissoort die ik juist had uitgekozen.
Maar Sjouk redde de situatie.
"Ik wil beide wel", zei ze rustig, "maar dan betalen Anne en ik de wijn wel bij het eten en het nagerecht". De dames gingen ogenblikkelijk akkoord. Hadden ze ooit kunnen bevroeden hoeveel plezier wij er later om gehad hebben, ze hadden spijt gehad als haren op hun hoofd dat ze ons niet het hele diner hadden laten betalen. Want wie er daarna ook op bezoek kwam bij ons, wij vroegen met stalen gezichten: "Wat willen jullie. Nu een kopje koffie, of vanavond een borrel?"