Werkplaats van het sociaal onderzoek

Maten voor de interne consistentie van de Guttmanschaal

P. G. Swanborn

Een bonte verzameling van maten voor de interne consistentie van Guttmanschalen wordt toegepast in onderzoeksrapportages. Nog steeds wordt hier en daar alleen de „rep" gerapporteerd; soms in combinatie met een „minimumrep"-maat (1), alsof de superioriteit van „kansrep"-maten niet reeds lang is aangetoond. Mokken (2) signaleert dit en geeft duidelijke richtlijnen; door de wiskundige verpakking is echter een aanzienlijk deel ervan voor de doorsnee geïnteresseerde ontoegankelijk. Het lijkt ons daarom zinvol enkele aspecten van deze kwestie weer te geven, waarbij we explicieter dan Mokken dit doet, ingaan op de parallelliteit van Greens I en Loevingers H. In recente publicaties worden nl. soms beide maten berekend, waarbij men Mokken volgt in diens vuistregels voor de H, maar, inconsequent de oude vuistregel van .50 voor de I blijft hanteren. In sectie 1 behandelen we de „normale" reproduceerbaarheidsmaten en Greens I; in sectie 2 gaan we in op Loevingers H, in sectie 3 worden ze met elkaar vergeleken. In sectie 4 geven we een berekeningsvoorbeeld, terwijl tenslotte in sectie 5 op dichotomiseringsproblemen wordt ingegaan.

1. De meest bekende maten zijn de reproduceerbaarheid (rep), de kansreproduceerbaarheid (kansrep) en Greens (3) Index of Consistency (I). Centraal staat de maat

$$\text{rep} = 1 - \frac{e}{Nk} \qquad \text{waarin } \begin{array}{l} e = \text{aantal fouten} \\ N = \text{aantal eenheden} \\ k = \text{aantal items} \end{array}$$

Niet alleen echter is de ondergrens van deze maat ongelijk aan nul omdat het maximale aantal fouten nooit meer dan ongeveer de helft van Nk kan zijn (4), ook zal duidelijk zijn dat we onze rep in ieder geval groter willen laten zijn dan de rep die verkregen wordt wanneer de items statistisch onafhankelijk zijn. Dit laatste is wel de minimumeis die we aan de itemset willen stellen. De kansrep wordt berekend op basis van de randfrequenties. Het afzetten van de rep tegen de kansrep gebeurt via de bekende formule

$$I = \frac{\text{rep} - \text{kansrep}}{1 - \text{kansrep}}\ ,$$ waarmee een maat voor de interne consistentie is gevonden die varieert tussen 0 en 1, en waarvan een bepaalde grootte als een noodzakelijke en voldoende voorwaarde kan worden gehanteerd voor de bruikbaarheid van de items als Guttmanschaal.

De eis dat de rep groter moet zijn dan .90 is overbodig (evenals alle andere vuistregels); het enige waar het op aan komt is de verhouding tussen het werkelijke aantal fouten e en het aantal toevalsfouten bij onafhankelijkheid e_{kans}. Bij omwerking blijkt nl.

$$I = \frac{(1 - \frac{e}{Nk}) - (1 - \frac{e_{\text{kans}}}{Nk})}{1 - (1 - \frac{e_{\text{kans}}}{Nk})} = \frac{\frac{e_{\text{kans}}}{Nk} - \frac{e}{Nk}}{\frac{e_{\text{kans}}}{Nk}} = 1 - \frac{e}{e_{\text{kans}}}\ ,$$

of m.a.w. I is gelijk aan 1 minus het aantal werkelijke fouten gedeeld door het aantal verwachte fouten bij onafhankelijkheid. Hieruit blijkt dat men rep en kansrep niet eens hoeft te berekenen. Het zal ook duidelijk zijn, dat het min of meer geheimzinnige begrip *„unidimensionaliteit"* niets anders betekent dan *associatie, homogeniteit* of *interne consistentie* van een set items. Men kan eventueel ook als alternatief het woord *betrouwbaarheid* gebruiken als men maar bedenkt dat het hier inter-item betrouwbaarheid betreft. De term *cumulatief* betekent in de empirie hetzelfde, plus het feit dat de items verschillende populariteiten hebben. De term *reproduceerbaarheid* heeft betrekking op de mogelijkheid (bij een perfecte schaal) dat van jantje, als hij score 4 heeft, precies het antwoordpatroon op de items gereproduceerd kan worden vanuit deze score. Nu is niemand hier in de praktijk in geïnteresseerd, en bovendien verliest dit feit aan betekenis wanneer (zoals altijd) de schaal niet perfect is. Vandaar, dat we het begrip reproduceerbaarheid (en daarmee de rep en de kansrep) beter kunnen vergeten.

Ons verhaal tot nu toe reduceert zich tot de maat voor de interne consistentie I, waarbij we aantekenen dat de gebruikelijke „.50 eis" volstrekt willekeurig, nogal stringent en te ongenuanceerd is. Mokken formuleert voor het geval van de H de volgende vuistregel:

$.30 \leqslant H < .40$: zwakke schaal
$.40 \leqslant H < .50$: matige schaal
$H \geqslant .50$: sterke schaal.

Er bestaat geen enkele reden om, wanneer men niet de H maar de I berekent, niet eén dergelijke vuistregel, die genuanceerder en minder stringent, maar helaas even willekeurig is, over te nemen. Dit te meer, waar de I en de H elkaar in de praktijk qua grootte weinig ontlopen. Het is onjuist om zowel een I als een H te berekenen, verschillende vuistregels te hanteren en bij niet voldoen aan één van de vuistregels ($I > .50$) de schaal te verwerpen. Om de gedachten te bepalen: ligt de I op .40, dan ligt de H meestal tussen .35 en .45, en vaak binnen aanzienlijk nauwere grenzen. Een en ander wordt acceptabeler wanneer we in sectie 2 de afleiding geven waaruit blijkt dat ook de H de vorm heeft van 1 minus het aantal fouten gedeeld door het aantal kansfouten (5); het verschil met de I zit alleen in de telling van het aantal fouten en het aantal kansfouten.

2. Loevinger (6) heeft reeds in 1948 enkele homogeniteitscoëfficiënten ontwikkeld, waarvan we noemen de h_{ij} (de homogeniteitscoëfficiënt over 2 items) en de H (de schaalhomogeniteitscoëfficiënt). De h_{ij} is gelijk aan φ / φ_{max}. De formules zijn als volgt:

$$(1)\quad h_{ij} = \frac{p_{ij} - p_i\, p_j}{p_i\,(1 - p_j)}$$

waarin $p_i < p_j$ (p_i is het „moeilijker" item) en p_{ij} is de vulling van, let wel, de ++ cel.

$$(2)\quad H = \frac{\sum\limits_{i=1}^{k-1} \sum\limits_{j=i+1} p_i\,(1 - p_j)\, h_{ij}}{\Sigma\,\Sigma\ p_i\,(1 - p_j)}$$

of H is het gewogen gemiddelde van alle item-paren-h's.

Invullen van (1) en (2) waarbij we ten gerieve van de drukker en de lezer de sub- en superscripts maar weglaten, levert:

$$H = \frac{\Sigma\,\Sigma\ p_i\,(1 - p_j)\, \dfrac{p_{ij} - p_i\, p_j}{p_i\,(1 - p_j)}}{\Sigma\,\Sigma\ p_i\,(1 - p_j)}$$

$$= \frac{\Sigma\,\Sigma\ (p_{ij} - p_i\, p_j)}{\Sigma\,\Sigma\ p_i\,(1 - p_j)}$$

$$= \frac{\Sigma\,\Sigma\ p_i\,(1 - p_j) - \Sigma\,\Sigma\ (p_i - p_{ij})}{\Sigma\,\Sigma\ p_i\,(1 - p_j)}$$

$$= 1 - \frac{\Sigma\,\Sigma\ (p_i - p_{ij})}{\Sigma\,\Sigma\ p_i\,(1 - p_j)}$$

$$= 1 - \frac{\Sigma\,\Sigma\ p_{e*}}{\Sigma\,\Sigma\ p_{e^*_{kans}}}$$

$$= 1 - \frac{\Sigma\,\Sigma\ e^*}{\Sigma\,\Sigma\ e^*_{kans}}$$, of, ter vereenvoudiging van de notatie,

$$H = 1 - \frac{e^*}{e^*_{kans}}$$

Merkwaardigerwijs heeft Greens I een veel grotere bekendheid gekregen dan Loevingers H. Een reden is waarschijnlijk, dat Greens maten dichter staan bij de oorspronkelijke Guttmanschaal-idee, door het behoud van op de reproduceerbaarheidsgedachte gebaseerde maten, dan de H-maten die Guttmans model onbarmhartig „entmythologisierten", door er een associatiematen-benadering op los te laten, iets waar Guttman niets voor voelde (8).

In Nederland heeft Mokken in zijn proefschrift en computerprogramma's de item- en schaalhomogeniteitscoëfficiënten centraal gesteld.

3. We zagen dat Greens I en Loevingers H dezelfde vorm hebben, maar verschillen in de wijze waarop de fouten worden geteld; hierdoor treden verschillen tussen beide maten op. Op deze foutentelling gaan we nu nader in.

Wanneer men de items in volgorde van afnemende moeilijkheidsgraad (dus toenemende populariteit) noteert, wordt voor de rep (waarop de I gebaseerd is) elke +- opeenvolging tussen items i en i + 1 als een fout geteld. Heeft men de patronen uitgeschreven, dan kan simpelweg geturfd worden, anders kunnen de betreffende kruistabellen tussen alle i/i + 1 itemparen uitsluitsel geven. Men mist op deze wijze de fouten van de tweede en hogere soorten, resp. van de vorm ++--, +++--- e.d. Fouten van de derde en hogere soort verwaarloost men altijd. Fouten van de tweede soort worden via visuele inspectie geturfd (bij uitgeschreven patronen), òf geteld uit de 4-ingangs kruistabellen van de items i, i + 1, i + 2, i + 3 (ten behoeve van repA), òf geschat uit de combinatie van de 2-ingangs kruistabellen van de items i/i+2 en i+1/i+3 (repB).

Voor de berekening van de kansrep werkt men eveneens, uiteraard, alleen met de kruistabellen van de items i/i+1 (fouten van de eerste soort) en i, i+1, i+2, i+3 (fouten van de tweede soort). Deze kansfouten worden berekend op basis van de randfrequenties.

Voor de H-berekening wordt daarentegen de feitelijke vulling en de vulling-bij-toeval van *alle* 1/2n(n-1) itemparen ij in de berekening betrokken. Mokken noemt dit een voordeel van de H boven de I. Zonder dat wij nu de I boven de H prefereren (welke maat men gebruikt is o.i. om het even) moeten we hier toch een vraagteken plaatsen. Het lijkt ons weinig zin te hebben om eenzelfde fout een aantal malen te tellen, zoals in de H-berekening gebeurt. Ons bezwaar richt zich niet tegen het feit dat een +- opeenvolging nu eens éénmaal, dan weer bijv. viermaal wordt geteld als fout, afhankelijk van de plaats in het antwoordpatroon (zie fig. 1).

a. +-+++

b. ++++-

Hier is geen bezwaar tegen omdat de randfrequenties doorwerken in de kansrep, waartegen de werkelijke fouten immers afgewogen worden. Veeleer missen wij het simpele idee ontwikkeld vanuit het cumulatieve patroon: elke keer dat een + gevolgd wordt door een -, is dat één fout. De foutentelling voor de I spreekt ons daarom wat meer aan. De belangrijkste conclusie is echter, dat I en H vergelijkbare maten zijn, en dat eventuele vuistregels voor de éne toegepast kunnen worden voor de andere.

4. Het lijkt zinvol om met een voorbeeld de consequenties van de verschillende maten te verduidelijken. In onderstaande onderdriehoek van een vierkante matrix zijn alle 2 × 2 kruistabellen opgenomen van een aantal items waarvan de schaalbaarheid onderzocht wordt. We gaan er van uit dat vrijwel altijd tijdens een schaalconstructiepoging een dergelijke matrix, rechtstreeks gebaseerd op de kruistabellen, gemaakt wordt. Berekening van de I maakt gebruik van de tabellen op de hoofddiagonaal en de diagonaal daaronder; berekening van de H van alle tabellen. Met een tafelrekenmachine

Figuur 1

	+ −							
+ 2 −	8 19 16 13 83	+						
+ 3 −	14 33 13 7 69	19 28 17 8 68	+					
+ 4 −	15 48 10 6 54	22 41 13 5 55	29 34 23 18 42	+				
+ 5 −	18 72 6 3 30	26 64 7 1 32	42 48 13 5 28	56 34 17 7 26	+			
+ 6 −	20 74 5 1 28	25 69 6 2 27	41 53 11 6 23	55 39 15 8 21	77 17 21 13 16	+		
+ 7 −	18 77 5 3 25	21 74 6 6 22	45 50 11 2 26	55 40 14 8 20	77 18 21 13 15	77 18 21 17 11	+	
+ 8 −	19 86 3 2 16	25 80 4 2 16	45 60 7 2 16	57 48 9 6 12	84 21 13 6 12	88 17 14 6 12	88 17 14 7 11	+
+ 9 −	20 91 2 1 11	26 85 3 1 11	44 67 5 3 9	62 49 6 1 11	88 23 9 2 10	89 22 9 5 7	89 22 9 6 6	99 12 10 6 6
	1	2	3	4	5	6	7	8

Toelichting bij figuur I

De omcirkelde getallen geven de verwachte vulling van de „nulcellen" bij onafhankelijkheid van de items.

De H-berekening verloopt als volgt:

$e = 13 + 7 + 6 + 3 \ldots\ldots = 208$

$e_{kans} = 16 + 13 + 10 + 6 \ldots = 388$

$$H = 1 - \frac{208}{388} = .46$$

De I-berekening verloopt als volgt:

$$e = 13 + 8 + 18 + 7 + 13 + 17 + 7 + 6 + \frac{7 \times 5 + 5 \times 5 + 5 \times 8 + 8 \times 13 + 13 \times 6 + 6 \times 6}{123}$$

$= 89 + 2.59 = 91.59$

$$e_{kans} = 16 + 17 + 23 + 17 + 21 + 21 + 14 + 10 + \frac{13 \times 13 + 13 \times 13 + 13 \times 15 + 15 \times 21 + 21 \times 14 + 14 \times 9}{123}$$

$= 139.7 + 10.31 = 150.01$

$$I = 1 - \frac{91.59}{150.01} = .39$$

Dit is tevens het grootste verschil dat we ooit tussen een I en een H vonden.

zijn beide maten in enkele minuten te berekenen. We stellen dit zo expliciet, om duidelijk te maken dat wanneer men over de kruistabellen (met een of andere associatiemaat, zoals de gamma) beschikt, het construeren van een Guttmanschaal zonder verdere computerhulp een kleinigheid is. We prefereren in sommige gevallen handwerk boven computerhulp, omdat de dysfunctie van het gebruik van computers in de sociologie, het onbegrip t.a.g. wat er eigenlijk gebeurt, nu en dan te evident wordt.

5. In de redactie van opinie- en attitudevragen is meestal gebruik gemaakt van 5 of meer antwoordcategorieën. Voor de dichotomisering kan men kiezen uit 2 oplossingen: òf alle items op dezelfde wijze splitsen (bijv. de positieve categorie wordt bij alle items gevormd door de antwoordcategorieën 1 + 2, of bij alle items door 1 + 2 + 3), òf men kan per item verschillend dichotomiseren en op basis van de straightruns een fraai „getrapte" reeks van populariteiten zoeken. Wij kozen tot nu toe altijd voor de laatste oplossing, en wel op grond van de volgende overweging: a) qua model is er geen bezwaar tegen; de antwoordcategorieën van alle items worden geacht ordinaal op het continuum te zijn geordend; b) naarmate het model beter wordt benaderd (hogere I of H) is de grootte van de populariteitsintervallen van toenemend belang i.v.m. de classificatie van de eenheden. Stel bijv. dat 5 items de volgende populariteiten hebben: 20%; 35%; 52%; 70% en 90% positief, dan betekent dit dat onder het perfecte model de populatie opgesplitst wordt in groepen van resp. 20% (alles positief); 15%; 17%; 18%; 20% en 10% (alles negatief). Hebben i en i + 1 populariteiten van resp. 30 en 32%, dan omvat één van de respondentenklassen slechts 2% van de populatie, waardoor de schaaldifferentiatie aan betekenis verliest.

Gaat men echter uit van schalen met een vrij kleine I of H, en scoort men de respondenten zoals gewoonlijk gebeurt, door simpelweg de plusjes te tellen, dan is het verband tussen de populariteitsintervallen en de klassenverdeling der respondenten veel minder sterk. Hierdoor is te verklaren dat sommige schalen soms tot 4 of 5 items met ongeveer gelijke populariteiten omvatten, maar toch tot een redelijke klassenverdeling van de respondenten leiden. Wanneer we de items uit figuur 1 op *gelijke* wijze dichotomiseren door overal de antwoordcategorieën 1 en 2 (van de 5) als positief te rekenen, ontstaat de volgende reeks populariteiten:

itemnr.	7	39%	*itemnr.*	9	90%
	8	41%		2	91%
	5	73%		3	92%
	6	76%		4	93%
	1	82%			

De verkregen schaal, met een I van .38 (H = .39) leidt bij „simple scoring" tot de volgende klassenverdeling der respondenten:

klasse	*frequentie*	*klasse*	*frequentie*
0	0	5	14
1	1	6	13
2	1	7	26
3	8	8	33
4	6	9	21

Combineert men de eerste 5 klassen, dan is dit een redelijke verdeling. Hadden we de items op categorie 1 versus „de rest" gedichotomiseerd, dan waren alle populariteiten gedaald, en zou de gevonden klassenverdeling naar de eerstgenoemde klassen verschoven zijn, tengevolge waarvan de laatste klassen vrij leeg zouden zijn. Het blijkt dus geenszins nodig te zijn om persé qua populariteit gesepareerde items in de schaal op te nemen. Wel is het wenselijk om zowel populaire als zeer weinig populaire items op te nemen, teneinde de kans op een redelijke vulling van alle klassen te vergroten. Tenslotte zij nog de klassenverdeling vermeld, die we verkregen op basis van de dichotomiseringen gebruikt in figuur 1:

klasse	*frequentie*	*klasse*	*frequentie*
0	3	5	18
1	4	6	27
2	9	7	25
3	10	8	9
4	12	9	6

Wat in feite een detailpunt is, zet ons nog eens aan het denken over de zinvolheid van het cumulatieve model, waarvan we in de praktijk wel ver af zijn geraakt. In feite is het hele systeem van de „simple scoring", het optellen van de positieve antwoorden, de doodgewone praktijk van een Likertachtige scoring. Wij willen geen pleidooi houden voor die andere wijze van scoring, waarbij eerst „imperfecte" antwoordpatronen worden omgevormd tot „perfecte" en vervolgens de score van het perfecte patroon krijgen toegekend; Mokken stelt terecht de onzinnigheid hiervan aan de kaak.

Maar de probabilisering van het model, en de verzwakking van de eisen (.30 i.p.v. .50) leiden wel tot de vraag in hoeverre het cumulatieve idee in het model, waarvan de werkelijkheid zo sterk afwijkt, nu zoveel meer inzicht of voordelen biedt dan een huis-, tuin- en keuken Likertschaal. Het valt echter buiten het kader van dit artikel om hier nader op in te gaan.

Noten

1. Ofshe, R. and L. Ofshe, „A comparative study of two scaling methods", in: *Sociometry* 1970, 33,4: 409-426.
2. Mokken, R. J., *A theory and procedure of scale analysis;* Mouton 1970, 59; 148-153; 182-194.
3. Green, B. F., A method of scalogram analysis using summary statistics; in: *Psychometrika,* 1956, 21; 79-88.
4. Galtung, J., *Theory and Methods of Social Research;* Universitetsforlaget Oslo 1967, 267-269.
5. Mokken geeft wel verschillende formules voor de H, maar niet de onderlinge afleidingen. Wij geven hier de afleiding vanuit de vorm die Loevinger in de oorspronkelijke publicatie (6) geeft.
6. Loevinger, J., The technique of homogeneous tests compared with some aspects of „scale analysis" and factor analysis; in: *Psychologisch Bulletin* 1948, 45, 507-530.
7. We noteren e* ter onderscheiding van de e in sectie 1.
8. Zie bijv. Guttman, L., The basis for scalogram analysis; in: Stouffer et al.: *Measurement and prediction,* Princeton Un. Press 1950, 60-90.