

Overeenstemming en associatie in sociologisch onderzoek

Roel Popping*

1. Inleiding

Hoewel overeenstemmingsmaten vooral ontwikkeld zijn binnen biologie, geneeskunde en psychologie, hebben enkele van de eerste geschriften die zich bezig houden met dit onderwerp, betrekking op een toepassing binnen de sociologie. Het betreft hier artikelen van Woodward en Franzen (1948) en van Scott (1955). Deze studies gebruiken overeenstemmingsmaten bij inhoudsanalyse om te achterhalen of stukken tekst betrouwbaar worden gecodeerd. Overeenstemmingsmaten worden hier gebruikt om na te gaan in hoeverre de inhoud van teksten door verschillende codeurs op dezelfde wijze is gecodeerd.

Gelet op de groeiende belangstelling voor kwalitatieve methoden van dataverzameling binnen de sociologie (naast inhoudsanalyse ook observatie), is het nodig ook meer aandacht te besteden aan overeenstemming of interbeoordelaarsbetrouwbaarheid. Dit is onontbeerlijk in de fase van een onderzoek waar de kwalitatief verkregen data gekwantificeerd moeten worden.

In paragraaf 2 wordt aangegeven wat het verschil is tussen associatie en overeenstemming. Vervolgens wordt in paragraaf 3 een overzicht van de ontwikkeling van overeenstemmingsmaten gegeven. In paragraaf 4 worden diverse situaties besproken waarin het binnen de sociologie zinniger is om overeenstemming te berekenen dan associatie.

In de daarop volgende paragrafen worden enkele uitbreidingen en problemen aangestipt, en er wordt kort ingegaan op de relatie met afstands- en gelijkenismaten.

* Sociologisch Instituut Rijksuniversiteit Groningen. De auteur dankt L. Dijkstra en F.N. Stokman voor hun waardevolle op- en aanmerkingen bij een eerdere versie van deze tekst.

2. Overeenstemming versus associatie

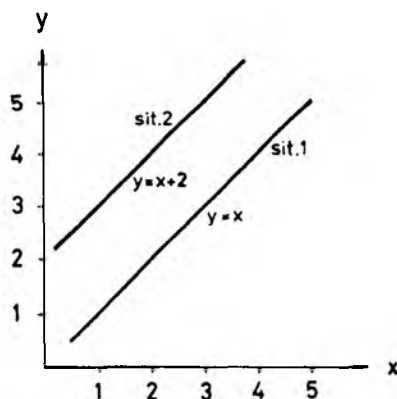
Bij overeenstemming gaat het om het vaststellen van de betrouwbaarheid, dat wil zeggen de consistentie tussen twee of meer classificaties van dezelfde data. Deze overeenstemming wordt uitgedrukt in de waarde van een index. Doorgaans betreft het data die gemeten zijn op nominaal niveau. Vaak gaat het om de classificaties door beoordelaars van dezelfde observaties; dan wordt gesproken over interbeoordelaarsbetrouwbaarheid. Het kan echter ook om andere vormen van classificatie gaan, zoals verderop zal worden getoond. Overeenstemming zegt niet direct iets over de mate waarin een kenmerk juist wordt weergegeven (validiteit).

Overeenstemming wordt vaak beschouwd als een speciale vorm van associatie. Er zijn echter verschillen. Dit kan worden verduidelijkt aan de hand van de onderzoeksdoelen en -problemen waarvoor ze geschikt zijn.

Bij associatie gaat het om gelijkenis naar de vorm tussen elementen die als variabelen kunnen worden opgevat (ook als het subjecten zijn), met als onderzoeksdoel het voorspellen van waarden op de ene variabele uit die op de andere, of omgekeerd. Hieronder vallen correlatie, covariantie, associatiematen op lagere meetniveaus (al dan niet genormeerd). Zij hebben een gemeenschappelijke grondslag welke bestaat uit conditioneel voorspellen.

Bij overeenstemming gaat het om de gelijkenis naar de inhoud van gedrag (in de meest ruime betekenis) tussen in het algemeen subjecten met als doelstelling het bepalen van de mate van identiteit van dit gedrag, zonder echter het gedrag van de een uit dat van de ander te willen voorspellen. Noodzakelijk bij overeenstemming is dat de tabel waarvan wordt uitgegaan, evenveel rijen als kolommen heeft, en dat de categorieën van de rijen in dezelfde volgorde staan als die van de kolommen. Corresponderende rijen en kolommen moeten naar dezelfde inhoud verwijzen. Twee rijen kunnen slechts verwisseld worden als dit ook gebeurt met de corresponderende kolommen. Twee observaties stemmen overeen als ze aan dezelfde categorie zijn toegewezen. Al deze observaties komen op de diagonaal van de overeenstemmingstabel terecht. Met betrekking tot de data in een tabel kan gelden dat een hoge associatie gevonden wordt, maar een lage overeenstemming. Het omgekeerde geldt niet. Dit wordt geïllustreerd in *grafiek 1*.

Voor situatie 1 in *grafiek 1* geldt dat er zowel perfecte associatie is als perfecte overeenstemming tussen X en Y, voor situatie 2 geldt ook dat er perfecte associatie is, maar de overeenstemming is hier verre van perfect. Perfecte associatie behoeft dus niet te betekenen perfecte overeenstemming, terwijl bij overeenstemming steeds gebruik moet zijn gemaakt van hetzelfde categorieënstelsel, bij associatie echter niet.



Zowel bij associatie als overeenstemming geldt met betrekking tot de categorieën dat deze elkaar uitsluiten en dat ze uitputtend zijn. Verder dat de beoordelaars onafhankelijk van elkaar werken.

Enkele Nederlandse studies waarin aandacht is geschonken aan betrouwbaarheid tussen beoordelaars worden behandeld in Muskens (1980).

3. Overeenstemmingsmaten en hun eigenschappen

3.1. Geschiedenis

Tegen de achtergrond van dit verschil tussen overeenstemming en associatie is het begrijpelijk dat in eerste instantie als maat voor overeenstemming het percentage gevallen is gebruikt waarin de beoordelaars het met elkaar eens zijn over het aan dezelfde categorie toewijzen van de onderzoekseenheden. Probleem hierbij is echter dat wanneer er weinig antwoordcategorieën zijn de kans op gelijke oordelen groter is dan bij veel antwoordcategorieën. Dit is het gevolg van het feit dat bij weinig antwoordcategorieën de totale kansruimte in minder partjes behoeft te worden verdeeld dan bij veel categorieën. Teneinde deze vertekening te voorkomen, hebben Bennett et al. (1954) hun 'coefficient of stability S' opgesteld, door Scott 'index of consistency S' genoemd, waarin rekening wordt gehouden met het aantal antwoordcategorieën. Bij deze grootheid wordt er van uitgegaan dat er twee beoordelaars zijn en dat voor beide beoordelaars afzonderlijk geldt dat elke antwoordcategorie een gelijke kans heeft om te worden gebruikt.

In werkelijkheid echter kan het zo zijn dat alle gevallen in slechts een paar categorieën worden ondergebracht. Dit maakt dat S geen bevredigende grootheid is. Scott heeft nu zijn 'coefficient of intercoder agreement', π_i , ontworpen waarin zowel gecorrigeerd wordt voor het aantal antwoordcategorieën als voor de mate waarin elk van deze categorieën wordt gebruikt (Scott, 1955). Deze maat is gedefinieerd als:

$$\pi_i = [P(o) - P(e)] / [1 - P(e)]$$

waar $P(o)$ de proportie oordelen weergeeft welke de beoordelaars onafhankelijk van elkaar aan dezelfde antwoordcategorieën hebben toegekend en $P(e)$ de proportie overeenstemming is volgens toeval. Deze laatste proportie hangt af van het aantal antwoordcategorieën en de mate waarin elke daarvan gebruikt is. Scott veronderstelt dat de verdeling van proporties over de categorieën bekend is voor de populatie en dat daarom voor elke antwoordcategorie geldt dat beide beoordelaars er steeds evenveel gevallen aan toewijzen. Hij stelt dat dit niet altijd precies zo zal zijn, maar dat het toch in het algemeen wel opgaat.

De toevalsproportie $P(e)$ is hier gedefinieerd als de som over alle categorieën van de kwadraten van de gemiddelde proportie toewijzingen per categorie.

3.2. Cohens kappa

Tegen de redenering van Scott heeft Cohen bezwaar gemaakt. Volgens hem zullen de proporties toewijzingen per categorie per beoordelaar meestal verschillen en slechts incidenteel even groot zijn: '. . . one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgements differently over the categories' (Cohen, 1960, p. 41). Vandaar dat hij $P(e)$ anders definieert dan Scott en wel als de som over alle categorieën van het produkt van de proporties per categorie van de beide beoordelaars. Zijn overeenstemmingsgrootheid noemt hij kappa:

$$\begin{aligned} \text{kappa} &= [P(o) - P(e)] / [1 - P(e)] \\ &= [F(o) - F(e)] / [N - F(e)] \end{aligned}$$

waarbij P proporties aangeeft, F frequenties en N het aantal te beoordelen gevallen. Zie ook *tabel 1*, waarin c het aantal antwoordcategorieën aanduidt.

Tabel 1. Voorbeeld van een overeenstemmingsmatrix; P_{ij} is de proportie die door beoordelaar 1 aan categorie i is toegewezen en door beoordelaar 2 aan categorie j

		beoordelaar 2				
		1	2	...	c	
beoordelaar 1	1	P_{11}	P_{12}		P_{1c}	$P_{1.}$
	2	P_{21}				$P_{2.}$
	...					$\vdots$
	i			P_{ij}		$\vdots$
	c	P_{c1}				$P_{c.}$
		$P_{.1}$	$P_{.2}$		$P_{.c}$	

$$P_o = \sum_{i=1}^c P_{ii}$$

$$P_e = \sum_{i=1}^c P_{i.} P_{.i}$$

Noem de proportie gevallen met betrekking waartoe geen overeenstemming is gevonden tussen de beoordelaars $Q(o)$ en de proportie waarvoor volgens toeval geen overeenstemming wordt verwacht $Q(e)$, dan is kappa eveneens te schrijven als:

$$\text{kappa} = [Q(e) - Q(o)] / Q(e)$$

omdat nu geldt: $Q(o) = 1 - P(o)$, en $Q(e) = 1 - P(e)$. Dit komt overeen met de verbale definitie die Cohen geeft: kappa is 'the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration' (Cohen, 1960, p. 40).

In zijn artikel geeft Cohen ook nog aan dat het gebruiken van de chi-kwadraattoets of de contingency coefficient C teneinde overeenstemming te meten, op geen enkele wijze te verdedigen is. Met behulp van deze grootheden wordt immers een nulhypothese getoetst die betrekking heeft op associatie (de mate waarin waarden op de ene variabele correct kunnen worden voorspeld uit die op de andere) en niet op overeenstemming: de mate van identiteit tussen het codeergedrag van twee beoordelaars.

Momenteel is Cohens kappa de meest gebruikte maat voor het berekenen van overeenstemming bij data die gemeten zijn op een nominale schaal.

3.3. Steekproefkenmerken

Voor kappa kunnen de steekproefkenmerken berekend worden. Er heeft echter veel discussie plaatsgevonden over de vraag hoe deze te berekenen (zie

bijv. Landis en Koch, 1975 voor een beschouwing daarover). Men moet zich ook afvragen waarvoor deze steekproefkenmerken gebruikt gaan worden. Is het om te onderzoeken of de gevonden uitkomst significant afwijkt van de uitkomst $\kappa = 0$, dan ontstaat er meteen een nieuw probleem. Vaak zal men namelijk (vooral als er veel observaties zijn) al bij vrij lage waarden van κ een niet-toevallige afwijking vinden van de hypothese ' $\kappa = 0$ '. Dit houdt in dat men in dat geval zijn twijfels moet hebben met betrekking tot de betrouwbaarheid tussen de beoordelaars. Door Landis en Koch (1977, p. 165) is gesteld dat κ groter moet zijn dan .60 om 'substantial' te zijn; boven .80 is de uitkomst 'almost perfect'. De uitkomst .80 wordt door Muskens (1980, p. 131) een 'convention of the trade' genoemd, een vuistregel.

Het is zinvol een betrouwbaarheidsinterval rondom de gevonden uitkomst te leggen en na te gaan of een minimaal vereiste uitkomst, bijv. deze .80, daar binnen valt.

4. Toepassingen binnen de sociologie

Hierboven hebben wij aangegeven dat overeenstemmingsmaten ontwikkeld zijn voor onderzoek naar interbeoordelaarsbetrouwbaarheid. Dit is ook de meest voorkomende toepassing. Er zijn echter zeer interessante andere toepassingsmogelijkheden. Deze zullen wij nu behandelen aan de hand van concrete voorbeelden uit sociologisch onderzoek.

Teneinde het voorgaande te verduidelijken wordt hierna van enkele onderzoeken dat deel besproken waarin gebruik is gemaakt van overeenstemmingsmaten. Waar nodig zal aangegeven worden waarom deze maten berekend moeten worden en geen associatiematen.

4.1. Voorbeeld Varga

Varga gebruikte overeenstemmingsmaten niet primair om interbeoordelaarsbetrouwbaarheid vast te stellen, maar om codeurs te selecteren. Dit gebeurt in het kader van een studie naar sociologische en socio-psychologische factoren welke van invloed zijn op het gebruik van chemische technologie in Hongarije. Een van de belangrijkste hypothesen die getoetst moest worden, is het bestaan van een positieve correlatie tussen het succes van transfer van een technologisch onderzoek en het aantal 'achievement oriented' personen in leidinggevende posities in die sector (Varga, 1972, p. 175).

Uit de literatuur komen drie leiderschapstypen naar voren, namelijk: 'the

achievement-', 'the power-' en 'the affiliation-orientated'. Teneinde deze drie typen personen te onderscheiden, is een test afgenomen, die gebaseerd is op de McClelland Achievement Motivation Test (McClelland, 1969). De personen kregen met behulp van een diaprojector vier plaatjes te zien, elk gedurende 20 seconden. Daarna moesten zij naar aanleiding van elk plaatje een kort opstel schrijven aan de hand van een aantal instructies.

Deze opstellen werden daarna gescoord op de drie leiderschapstypen. Ze konden steeds worden toegewezen aan een uit 15 mogelijke categorieën. De opstellen werden dus driemaal geclassificeerd. Voor één persoon zou dit een te omvangrijke taak zijn. Dit probleem is opgelost door te beginnen met een groep beoordelaars, daar tijdens een training de beste uit te selecteren, en deze beste de uiteindelijke codeerwerkzaamheden te laten verrichten. Op deze wijze kan worden gegarandeerd dat er uiteindelijk hoogstens zeer kleine verschillen zouden bestaan tussen de diverse beoordelaars.

Er waren twaalf beoordelaars. Deze moesten binnen het kader van hun training eerst de opstellen van 30 proefpersonen uit de publicatie van McClelland aan de categorieën toewijzen. In deze publikatie waren ook classificaties gegeven. Nu kon steeds de overeenstemming gemeten worden tussen de experts van McClelland en de beoordelaars. Dit is gebeurd met behulp van de pi van Scott. Nadat dertig opstellen waren geclassificeerd, bleek de overeenstemming tussen de beste beoordelaars en de experts te zijn: $\pi = .81$.

De volgende trainingsfase was het classificeren van de opstellen uit een eigen vooronderzoek. Hier zijn ook de beoordelaars onderling vergeleken. Ook is gelet op de mate van overeenstemming tussen de beoordelaars voor elk van de 15 categorieën. Dit om te achterhalen of er beoordelaars waren die niet helemaal zeker waren met betrekking tot een bepaalde antwoordcategorie.

Tenslotte werden vier beoordelaars geselecteerd op basis van deze training om de gegevens uit het eigenlijke onderzoek te coderen. De opstellen zijn hier eenmaal beoordeeld op elk van de drie doeleinden. Volgens Varga maakt het nu niet uit door welke beoordelaar dit werd gedaan, het resultaat zou toch vrijwel hetzelfde zijn.

In dit onderzoek zijn derhalve overeenstemmingsmaten gebruikt om beoordelaars te selecteren vanuit de onderzoeksdoelstelling van zowel een grote onderlinge identiteit van hun codeergedrag, als een zo groot mogelijke overeenstemming van dit gedrag met de opvattingen van experts, en niet om een hypothese te toetsen over de wederkerige voorspelbaarheid van het codeergedrag van beoordelaars. Een nieuwe, op kappa gebaseerde, methode voor het selecteren van de beste beoordelaar uit een groep van beoordelaars

op basis van optimale overeenstemming is gepresenteerd door Schouten (1982).

4.2. Voorbeeld Gadourek

In het tweede voorbeeld wordt aangegeven hoe overeenstemmingsmaten kunnen worden gebruikt om na te gaan of twee ongelijke indicatoren voor een kenmerk identiek zijn, alsmede om de overeenstemming tussen objecten met behulp van zo'n kenmerk te analyseren.

In zijn onderzoek naar sociaal culturele veranderingen in Nederland houdt Gadourek zich onder andere bezig met het vraagstuk van verandering van kerkgenootschap en ontkerkelijkheid (Gadourek, 1982, p. 70 e.v.). Hij gaat voor respondenten uit een steekproef uit 1975 na in hoeverre zij lid zijn van hetzelfde kerkgenootschap als hun ouders. Hiervoor is het nodig dat er een score gevonden wordt voor het kerkgenootschap van de ouders. Daartoe wordt eerst de kerkelijke oriëntatie van de vaders en de moeders vergeleken (zie tabel 2).

Voor deze gegevens wordt gevonden: $\kappa = .81$. Deze uitkomst is 'almost perfect', en kan zo geïnterpreteerd worden dat het er niet toe doet welke van de twee variabelen: kerkgenootschap waarvan moeder of dat waarvan vader lid, gebruikt wordt als variabele 'kerkgenootschap van de ouders'; de verschillen zijn marginaal.

Ook hier is de onderzoeksdoelstelling dus niet het toetsen van een hypothese over samenhang: de vraag of er een systematisch verband bestaat tussen de kerkelijke oriëntatie binnen 1 358 ouderparen, doch de vraag naar de mate van identiteit van twee indicatoren.

Nu kan de relatie onderzocht worden tussen het kerkgenootschap waarvan de respondent lid is en dat waarvan de ouders lid zijn. Dit resulteert in tabel 3 (we gaan uit van het kerkgenootschap van de moeder als dat van de ouders).

Tabel 2. Verdeling over kerkgenootschap van moeder en vader van de respondent

		kerkgenootschap waarvan vader lid			
kerkg. waarvan					
moeder lid					
geen	260	r-k	prot.	rest	totaal
r-k	24	7	17	5	289
prot.	70	456	19	5	504
rest	11	6	394	11	481
totaal	365	4	4	65	84
		473	434	86	1 358

Tabel 3. Verdeling over kerkgenootschap van de respondent en dat van diens ouders (geïndiceerd door kerkgenootschap moeder)

kerkg. waarvan moeder lid	kerkgenootschap waarvan respondent lid				totaal
	geen	r-k	prot.	rest	
geen	256	8	18	7	289
r-k	118	376	6	4	504
prot.	177	6	289	9	481
rest	49	6	6	23	84
totaal	600	396	319	43	1 358

Hier wordt gevonden: $\kappa = .57$. Zouden we ons op de kerkelijke oriëntatie van de vaders gebaseerd hebben, dan resulteert: $\kappa = .60$. De interpretatie hiervan moet zijn dat er verschillen bestaan: er is immers geen perfecte overeenstemming tussen het kerkgenootschap van de respondent en dat van zijn ouders. Uitgaande van de hypothese ' $\kappa = 0$ ', blijken de standaardafwijkingen voor zowel de gegevens uit *tabel 2* als *tabel 3* .02 te bedragen. Bij een verwerpingsgebied van 5 procent zijn de gevonden κ 's significant; die voor de relatie tussen het kerkgenootschap van de ouders en dat van de respondent is echter niet erg hoog. In plaats van de hypothese ' $\kappa = 0$ ' te toetsen voor de gegevens uit *tabel 3*, heeft het meer zin uit te gaan van de hypothese ' $\kappa = .80$ ', deze situatie is immers 'almost perfect'. De standaardafwijking behorend bij de gevonden κ van .57 bedraagt .02. Nu wordt als z -waarde gevonden $z = 13.89$; dit is bij de gekozen α significant. De gevonden κ is dus systematisch lager dan .80, met andere woorden: er zijn verschuivingen opgetreden wat betreft het kerkgenootschap waar de kinderen lid van zijn, vergeleken met dat van hun ouders.

Hier is de overeenstemmingsmaat gebruikt om een hypothese te toetsen die betrekking heeft op een verandering qua kerkelijke oriëntatie van ouders naar kinderen. De vraag die Gadourek stelt, is of de ondervraagden nog dezelfde oriëntatie hebben als hun ouders. Deze vraagstelling is een geheel andere dan die naar de voorspelbaarheid van de oriëntatie van het kind uit die van de ouders. In het laatste geval zou het toepassen van een associatiemaat in plaats van een overeenstemmingsmaat voor de hand hebben gelegen; het gaat dan immers om de vraag naar het bestaan van een systematisch verband.

4.3. Voorbeeld De Vos

Het derde voorbeeld laat zien hoe overeenstemming gebruikt kan worden

voor het bepalen van de kappa tussen indelingen op basis van diverse schaal-analyses.

In zijn proefschrift vergelijkt De Vos een aantal technieken voor het meten van houdingen met elkaar (De Vos, 1980). Hij doet dit aan de hand van gegevens uit een onderzoek waarin men onder andere wilde nagaan in hoeverre Nederlandse arbeiders en ambtenaren de volgende vier oriëntaties met betrekking tot hun werk hebben: een instrumentele, een solidaristische, een bureaucratische en een professioneel-ambachtelijke oriëntatie.

Teneinde de oriëntaties te kunnen meten, zijn aan de respondenten 44 attitude-uitspraken voorgelegd. De Vos heeft een tiental schaaltechnieken gebruikt om te onderzoeken of de vier typen oriëntaties inderdaad voorkomen. Met behulp van vier technieken was het mogelijk te komen tot vier schalen, die elk een van de oriëntatie-typen aanduiden. De vier schaalmethoden zijn:

1. principale componenten analyse gevolgd door een varimax rotatie. De totale scores zijn verkregen door ongewogen optelling van items;
2. als 1, de totale scores zijn echter verkregen door optelling van de factor-scores;
3. de uitspraken zijn van te voren ingedeeld in vier groepen op grond van de oriëntatie die ze verwacht worden te meten. Elke groep is geanalyseerd door middel van principale componentenanalyse zonder rotatie. De totale scores zijn verkregen door ongewogen optelling van items;
4. als 3, per groep is echter een schaal geconstrueerd op grond van de item-restcorrelaties en Cronbachs alpha. Via ongewogen sommatie zijn de totale scores verkregen.

De respondenten zijn toegewezen aan een van de oriëntaties op grond van hun hoogste score op een van deze vier schalen. Hiervoor was het nodig dat eerst de resultaten die verkregen waren via de methodes 1, 3 en 4 gestandaardiseerd werden. Onderzocht is vervolgens in hoeverre de classificaties van de respondenten via de vier genoemde methoden met elkaar overeenstemmen. Er is dus niet nagegaan hoe hoog de vier schalen onderling correleren, maar onderzocht is in hoeverre de respondenten hetzelfde type oriëntatie hebben. Hiertoe is voor elk paar van classificaties op grond van de resultaten van de schaalmethoden Cohens kappa berekend (zie tabel 4):

Deze resultaten geven aan dat de methoden nogal verschillende uitkomsten opleveren; alleen de classificaties via de methoden 3 en 4 vertonen een dermate grote overeenstemming dat geconcludeerd mag worden dat het bij een volgend onderzoek niet veel uitmaakt welke van deze twee methoden gebruikt wordt. Let echter wel, er wordt hiermee niets beweerd over betrouwbaarheid en validiteit van de schalen. Het feit dat de uitspraken van te voren

Tabel 4. *Kappa's voor vier door De Vos gebruikte methoden*

methode 2	.59		
methode 3	.61	.50	
methode 4	.48	.58	.79
	methode 1	methode 2	methode 3

in vier groepen zijn ingedeeld, kan bepalend zijn geweest voor de hoge uitkomst.

De Vos heeft ook voor alle gevallen het verband berekend op grond van de associatiemaat Cramers V. Hij vindt steeds waarden die hoger zijn dan of gelijk aan kappa. Met Cramers V wordt echter niet de overeenstemming tussen twee classificaties berekend, maar het verband daartussen. Daarbij zijn ook de cellen betrokken die naar verschillende oriëntaties verwijzen.

De vraag kan gesteld worden of kappa hier wel gebruikt had mogen worden, immers een van de assumpties voor het gebruik is dat de observaties onafhankelijk van elkaar zijn en dat is hier niet het geval. De gehanteerde schalen zijn vaak opgebouwd uit dezelfde items. Strikt genomen mag kappa hier dus niet worden toegepast.

Iets dat vergelijkbaar is met wat De Vos heeft gedaan, is te vinden bij Wiedemann en Fenster (1978), die bij het interpreteren van de uitkomsten verkregen via discriminantanalyse ook gebruik maken van overeenstemmingsmaten.

4.4. *Voorbeeld Killworth en Bernard*

Tenslotte een voorbeeld uit een onderzoek waarin geen gebruik is gemaakt van overeenstemmingsmaten, terwijl dit wel te verkiezen zou zijn geweest. Overeenstemming kan hier worden gebruikt om na te gaan of indelingen op grond van cognities identiek zijn aan die op grond van feitelijk gedrag. Dit laatste kan hierbij als een criterium gelden.

In hun studie rapporteren Killworth en Bernard over een onderzoek naar triadestructuren die binnen groepen voorkomen (Killworth en Bernard, 1979/80). Een van de datasets die ze gebruikt hebben, is afkomstig van 40 personen die op een kantoor werken. Deze personen vormen 9 880 triaden. Voor elk van deze triaden is nagegaan hoe de contacten daarbinnen liggen en hoe de personen die tot een triade behoren, zeggen dat deze contacten zijn. Er zijn dus gegevens verzameld met betrekking tot een gedrags- en een cognitieve component. In totaal zijn er 16 typen triaden mogelijk. Elke tria-

de is tweemaal aan een van deze typen toegekend, eenmaal op grond van de gedrags- en eenmaal op grond van de cognitieve component. Killworth en Bernard willen onderzoeken in hoeverre de beide componenten overeenstemmen. Ze gaan daarbij uit van de 16×16 -matrix welke de frequenties bevat.

Het eerste wat ze doen, is deze frequenties correleren op grond van rangnummers 1 tot en met 16. Ze merken op dat dit niet helemaal bevredigend is, want 'it relies on subjective decisions as to which cognitive triads are more inaccurate than others when compared to a given behavioral triad (Killworth en Bernard, 1979/80, p. 37). Ze vinden een correlatie van .46.

Vervolgens passen ze een tweede 'measure of accuracy' toe. Per type gedragstriade wordt nu de proportie gevallen berekend, die tot hetzelfde type cognitieve triade behoort. De totale proportie gevallen die correct weergegeven is, bedraagt .45. Dit is de proportie gevallen op de diagonaal. Het type triade, waarin geen enkel contact tussen de drie personen voorkomt, is het vaakst juist weergegeven, hier is de proportie .58.

In dit onderzoek is geen overeenstemmingsmaat berekend, terwijl dat juist hier op grond van het onderzoeksdoel zeer voor de hand had gelegen. Nu zijn zelfs correlaties berekend, waarvan sterk moet worden betwijfeld of deze wel het kenmerk uitdrukken waar het hier om gaat. Kappa berekend op grond van de gegevens in de 16×16 -matrix geeft een waarde van .23, met andere woorden de overeenstemming tussen gedrags- en cognitieve structuur is erg laag. Bij het berekenen van hun tweede maat hebben Killworth & Bernard de invloed van de onafhankelijke variabele gedragstriade na willen gaan op de afhankelijke variabele cognitieve triade. De meest voor de hand liggende maat is hier Goodman en Kruskals lambda, deze maat heeft echter het nadeel dat ze voorbij gaat aan de inhoud van de afzonderlijke categorieën. Bovendien geeft ze slechts een uitkomst, terwijl de onderzoekers geïnteresseerd waren in de overeenstemming bij de afzonderlijke triaden. Voor dat geval zou gebruik moeten worden gemaakt van de conditionele kappa, die behandeld wordt in paragraaf 7.1 (de waarde die voor lambda wordt gevonden, bedraagt .04).

5. Betrouwbaarheid van beoordelaars versus beoordelingen

Vaak worden overeenstemmingsmaten slechts beschouwd als maten voor interbeoordelaarsbetrouwbaarheid. Het gaat er dan om te onderzoeken in hoeverre twee of meer beoordelaars de eenheden op dezelfde wijze classificeren. In de inleiding werd deze toepassingsmogelijkheid genoemd.

Er kan zich hier een probleem voordoen wanneer niet alle eenheden ge-classificeerd zijn door dezelfde beoordelaars. Stel dat tien objecten als volgt tweemaal zijn geclassificeerd door vijf beoordelaars (aangeduid met A tot E):

object	1	beoordeeld door	A en C,
..	2	..	.. A .. C,
..	3	..	.. A .. D,
..	4	..	.. A .. D,
..	5	..	.. B .. D,
..	6	..	.. B .. D,
..	7	..	.. B .. E,
..	8	..	.. B .. E,
..	9	..	.. C .. E,
..	10	..	.. C .. E.

Men kan in dit geval de overeenstemming tussen de beoordelaars onmogelijk bepalen, omdat daarvoor nodig is dat elk object is geclassificeerd door elk der beoordelaars. Dat is hier niet het geval. Cohens kappa kan hier dus niet worden gebruikt, ook niet om de overeenstemming te bepalen tussen de twee beoordelingen of classificaties. Immers de vijf beoordelaars kunnen elk verschillende criteria hebben gebruikt. Fleiss (1971) heeft daarom een uitbreiding van kappa voorgesteld, die bedoeld is voor de situatie waarin niet alle eenheden door dezelfde beoordelaars zijn geclassificeerd. Deze index geeft informatie over de betrouwbaarheid van de beoordelingen en niet over de betrouwbaarheid tussen de beoordelaars.

6. Enkele problemen

Bij het classificeren van de eenheden wordt de interbeoordelaarsbetrouwbaarheid bepaald door een combinatie van factoren. Hiertoe behoren:

- de categorieën; zijn deze goed gedefinieerd, zijn ze ruim of eng, complex of niet? Het categorieënstelsel dient uitputtend te zijn en de categorieën moeten elkaar wederzijds uitsluiten;
- de onderzoekseenheden die geclassificeerd moeten worden; zijn deze duidelijk herkenbaar? Met betrekking tot dit punt kunnen er nog extra problemen ontstaan als deze eenheden gevormd worden door tijdsintervallen, waarbij aangegeven moet worden of een bepaald gedrag al dan niet heeft plaatsgevonden binnen een interval. In welk tijdsinterval moet een gedrag

gescoord worden dat plaatsheeft tijdens de overgang van het ene interval in het andere? Een probleem dat hier ook speelt, is dat als een beoordelaar maar even wordt afgeleid het al mogelijk is dat hij een interval een of meer seconden langer laat duren dan is toegestaan. Alle latere tijdsintervallen zullen dan ook opschuiven. Er is veel literatuur over deze problemen (Mitchell, 1979) en ook zijn er technische hulpmiddelen ontwikkeld om het coderingsproces te vergemakkelijken.

– allerlei vormen van vertekening. Dit kan aan de beoordelaars liggen, hier komen we later op terug. Het kan ook aan de te beoordelen eenheid liggen, bijvoorbeeld wanneer dit een persoon is; deze kan zich tegenover de beoordelaars anders gedragen dan normaal.

Tot besluit van deze paragraaf zullen we ingaan op de vraag hoe te handelen als de waarde voor de overeenstemmingsmaat gevonden is, en vooral wat te doen als de overeenstemming laag is. Veel onnodige narigheid kan voorkomen worden. Er moet gewerkt worden aan de hand van duidelijke definities, liefst op papier gesteld. Beoordelaars dienen goed geïnstrueerd te zijn, ze moeten precies weten wat er van hen verwacht wordt en hoe ze het probleem moeten benaderen. Vaak is een training wenselijk. Ook is het aan te raden een aantal beoordelingen vooraf uit te laten voeren, waarbij het doel vooral is het selecteren van beoordelaars. Zij die de opdracht slecht uitvoeren, kunnen dan voor het eigenlijke onderzoek al verwijderd worden. Dit is wat Var-ga bij zijn onderzoek heeft gedaan.

Blijkt de gevonden overeenstemming ondanks alle voorzorgen toch laag te zijn, dan moet eerst gezocht worden naar oorzaken daarvan. Vaak zal dan een van de hiervoor genoemde punten gevonden worden: een beoordelaar functioneert niet goed, of de opdracht was niet duidelijk. Afhankelijk van het onderzoek waarbinnen de beoordelingswerkzaamheden hebben plaatsgehad, zijn er nu twee mogelijkheden. De eerste is dat alle beoordelingen opnieuw worden gedaan met betere instructies en/of nieuwe beoordelaars, de tweede is dat op dit punt het onderzoek (of het betreffende deel er uit) wordt afgesloten.

Blijkt de gevonden overeenstemming niet bijzonder groot te zijn, bijvoorbeeld tussen .40 en .80, dan wordt vaak verder gegaan met de score die het meest voorkomt, of, als dit mogelijk is, met de mediaan of het gemiddelde. Hierbij is het wel nodig dat de onderzoeker de nodige voorzichtigheid in acht neemt. Vaak is het evenwel wenselijk ook hier de beoordelingen te laten overdoen. Worden er weer geen hoge uitkomsten gevonden, dan mag dit beschouwd worden als een ondersteuning voor het besluit verder te gaan met de meest voorkomende score, de mediaan of het gemiddelde.

Als er een hoge waarde wordt gevonden voor de overeenstemmingsmaat, dan is het toch wenselijk dat de onderzoeker zich afvraagt of daar wellicht een irrelevante reden voor is. Dit is vooral van belang als er om bepaalde redenen juist geen hoge mate van overeenstemming werd verwacht. Bij een hoge overeenstemming maakt het nauwelijks uit met welke classificaties verder wordt gewerkt. Zie nogmaals het behandelde voorbeeld uit het onderzoek van Gadourek.

7. Uitbreidingen van de kappa-maat

De kappa-maat is op diverse wijzen uitgebreid, waardoor hij in uiteenlopende situaties bruikbaar is. Hierna zal op één ervan uitvoerig worden ingegaan. Deze uitbreiding kan van belang zijn binnen sociologische onderzoeken. De overige uitbreidingen zullen kort aangestipt worden.

7.1. *Condities*

Het kan zijn dat een onderzoeker niet tevreden is met alleen de mate van overeenstemming tussen de twee beoordelaars of vergelijkingseenheden.

Doorgaande op het voorbeeld van De Vos: de respondenten uit het oorspronkelijke onderzoek zijn getrokken uit bedrijven waarvan verondersteld werd dat de verwachte oriëntaties, als ze zouden voorkomen, daar dan zeker zouden voorkomen. De bedrijven vertegenwoordigen de eerste drie verwachte oriëntaties. Nu kan nagegaan worden in hoeverre de bedoelde oriëntatie voorkomt bij het bedrijfstype.

Passen we dit toe op de classificaties op grond van de als eerste genoemde methode van schaalanalyse, dan vinden we:

gegeven: verwacht wordt een bedrijf

met instrumentele werknemers $\text{kappa} = .12$

met solidaristische werknemers $\text{kappa} = .12$

met bureaucratische werknemers $\text{kappa} = .01$

De formule om deze conditionele kappa te berekenen, is geleverd door Light (1971, p. 367-368). Het blijkt dat de verwachting die gegolden heeft bij het trekken van de steekproef niet opgaat (Popping, 1981).

Met behulp van deze werkwijze is het mogelijk verdergaande conclusies te trekken, dan alleen maar het vaststellen van de mate van overeenstemming. In het voorbeeld hier is aangegeven dat de verwachtingen voor geen enkel bedrijfstype opgaan.

In hun onderzoek hadden Killworth en Bernard ook met vrucht gebruik

kunnen maken van deze conditionele kappa. Ze zouden dan gevonden hebben dat de hoogste overeenstemming niet gevonden wordt bij de triade zonder contacten, maar bij de triade waarin alle mogelijke contacten daadwerkelijk voorkomen. Voor deze triade wordt gevonden $\kappa = .46$, terwijl voor de eerste geldt dat $\kappa = .30$ is. Tussen deze twee triaden zit nog een ander type triade, namelijk dat waarin twee personen wederzijds contact hebben, maar geen contact met de derde persoon. Daarvoor geldt $\kappa = .32$.

7.2. Andere uitbreidingen

Andere uitbreidingen zijn dat de onderzoeker kan verdisconteren hoe ernstig hij het vindt dat een onderzoekseenheid aan verschillende antwoordcategorieën is toegewezen; dit gebeurt door wegingsfactoren in te voeren. Bij het berekenen van kappa worden per cel de geobserveerde en de verwachte hoeveelheid eenheden vermenigvuldigd met een bepaalde waarde. Deze waarde bedraagt 1 als de eenheid in beide gevallen aan dezelfde categorie is toegewezen, is dit niet zo, dan is deze waarde een getal tussen 0 en 1. De waarden vormen eigenlijk een schaal aflopend naar gelang de ernst van het niet-overeenstemmen. Op deze manier kan kappa ook geschikt worden gemaakt voor het bepalen van de overeenstemming bij data die gemeten zijn op ordinaal niveau.

Vervolgens kan kappa behalve voor twee, ook voor meer dan twee beoordelaars berekend worden. Hierbij zijn er twee mogelijkheden met betrekking tot de vaststelling van overeenstemming. De classificaties door de beoordelaars kunnen paarsgewijs vergeleken worden, maar ook kan een onderzoeker stellen dat er pas sprake is van overeenstemming als alle beoordelaars het eens zijn over het toewijzen van een object aan een categorie. Ook hier is het weer mogelijk kappa te berekenen per categorie en ook kan weer met gewichten gewerkt worden. Ook kan men een van de beoordelaars als de standaard beschouwen, dus er van uitgaan dat de toewijzingen van eenheden aan categorieën door hem volledig correct zijn gedaan. Tenslotte zijn er nog aanpassingen van de kappa-maat voor het geval er ontbrekende waarden zijn.

Om een heel ander type overeenstemming gaat het, waar de beoordelaars niet alleen de observaties aan de categorieën moeten toewijzen, maar ook deze categorieën nog moeten opstellen. Zie Popping (1983a) voor de behandeling van dit probleem.

8. Relatie met afstands- en gelijkenismaten voor subjecten

Everitt geeft een beschrijving van een onderzoek waarin beoordeeld moest worden op het wel of niet aanwezig zijn van een kenmerk. De beoordelaars waren een psychiater die de diagnose vanuit acht verschillende criteria stelde, het hospitaal en twee computerprogramma's. De kappa's voor alle paren van deze 11 beoordelingen van een medische diagnose, dienen als uitgangspunt voor een principale coördinatenanalyse (Everitt, 1978, p. 21-23). Deze methode verschilt van een principale componentenanalyse daarin dat de afstanden tussen de punten ook niet-euclidische afstanden tussen de oorspronkelijke observaties mogen weergeven. De kappa's worden hier beschouwd als gelijkenismaten, niet helemaal terecht overigens want Everitt geeft in zijn definitie van een gelijkenismaat aan dat deze waarden moet hebben die binnen het bereik 0-1 liggen, terwijl een aantal van de kappa's die we bij hem aantreffen negatief zijn.

In dit onderzoek zijn de kappa's berekend als maten van overeenstemming in het classificatiegedrag van twee beoordelaars. Daarna zijn ze als gelijkenismaat gebruikt in een analyse met als doel zicht te krijgen op de relaties tussen de verschillende diagnostische criteria.

Door Fleiss is aangetoond dat er veel gelijkenismaten zijn voor gegevens, die betrekking hebben op het wel of niet aanwezig zijn van een kenmerk, welke equivalent zijn aan de kappa-coëfficiënt zoals voorgesteld door Cohen (Fleiss, 1975).

Bij gelijkenismaten gaat het meestal om wel/niet-data. Een van de meest gebruikte van deze maten is de ruwe proportie-overeenstemming. Wil men deze maat gebruiken, dan moeten de frequenties wel en niet voorkomen van het kenmerk liefst ongeveer even groot zijn. Is dit niet zo, dan is een maat te prefereren waarin alleen overeenstemming gemeten wordt met betrekking tot die eenheden waarbij volgens tenminste een van de beoordelaars het kenmerk wel voorkomt, zoals bijv. de coëfficiënt van Jaccard. Bij gelijkenismaten wordt doorgaans niet gecorrigeerd voor toeval. Verder is het bij deze maten niet mogelijk de overeenstemming binnen een 'pool' van beoordelaars te berekenen; het gaat steeds om de gelijkenis tussen twee beoordelaars.

Bij gelijkenismaten wordt het aantal relevante matches tussen de twee beoordelaars over alle eenheden geteld. Het grote verschil met de afstandsmaten is dat bij deze laatsten ook de vereiste van de driehoeksongelijkheid geldt. Dit maakt ook dat er een wezenlijk verschil is tussen overeenstemmings- en afstandsmaten. Een uitgebreid overzicht van gelijkenis- en afstandsmaten is gegeven door onder andere Sneath en Sokal (1973).

In het voorgaande is er vrijwel steeds van uitgegaan dat de data met betrekking waartoe de overeenstemming moet worden vastgesteld, gemeten zijn op een nominale schaal. Bij hogere meetniveaus wordt vaak gebruik gemaakt van de intraclasscorrelatiecoëfficiënt en van conventionele psychometrische tests.

Een overzicht van wat er zoal beschikbaar is aan maten is gegeven door Landis en Koch (1975). Een recentere bibliografie van geschriften over overeenstemming verscheen van de hand van de auteur (Popping, 1979). Een groot aantal overeenstemmingsmaten voor nominale data kunnen berekend worden met behulp van computerprogramma's die voor dit doel zijn ontworpen. Een overzicht is te vinden in Popping (1983b).

Literatuur

- Bennett, E.M., Blomquist, R.L. en Goldstein, A.C., 'Communication through limited response questioning', *Public Opinion Quarterly*, 18, 1954, p. 303-308.
- Cohen, J., 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, 20, 1960, p. 37-46.
- Everitt B.S., '*Graphical techniques for multivariate data*', Londen, Heinemann, 1978.
- Fleiss, J.L., 'Measuring nominal scale agreement among many raters', *Psychological Bulletin*, 76, 1971, p. 378-382.
- Fleiss, J.L., 'Measuring agreement between two judges on the presence or absence of a trait', *Biometrics*, 31, 1975, p. 651-659.
- Gadourek, I., *Social change as a redefinition of roles. A study of structural and causal relationships in the Netherlands in the seventies*, Assen, Van Gorcum, 1982.
- Killworth, P.D. en Bernard, H.R., 'Informant accuracy in social network data III: A comparison of triadic structure in behavioral and cognitive data', *Social Networks*, 2, 1979/1980, p. 10-46.
- Landis, J.R. en Koch, G.G., 'A review of statistical methods in the analysis of data arising from observer reliability studies', *Statistica Neerlandica*, 29, 1975, p. 101-123, p. 151-161.
- Landis, J.R. en Koch, G.G., 'The measurement of observer agreement for categorical data', *Biometrics*, 33, 1977, p. 159-174.
- McClelland, D.C., 'The role of achievement orientation in the transfer of technology', in: Gruber, W.H. en Marguis, D.G. (red.), *Factors in the transfer of technology*, Cambridge (Mass.), The MIT Press, 1969, p. 61-81.
- Mitchell, S.K., 'Interobserver agreement, reliability, and generalizability of data collected in observational studies', *Psychological Bulletin*, 86, 1979, p. 376-390.
- Muskens, G.J., '*Frames of meaning – Are the measurable? A methodological critique of the content analysis of illustrated periodical magazines*', Nijmegen, 1980, dissertatie.
- Popping, R., 'Bibliography on measures of agreement', *Methoden en Data Nieuwbief*, 4, 1979, 2, p. 149-159.
- Popping, R., 'Een kleine kritiek op De Vos', in: Schuur, W.H. van, Popping, R., Rutges, A.C. en Stokman, F.N., Discussie over H.H. de Vos, 'Het meten van werkoriëntaties', *Bulletin no.*

- 42, Vakgroep Methoden en Technieken, Sociologisch Instituut, Groningen, 1981, p. 13-22.
- Popping, R., 'Traces of agreement. On the dot-product as a coefficient of agreement', *Quality and Quantity*, 17, 1983a, p. 1-18.
- Popping, R., 'Computer programs for nominal scale analysis', *Kwantitatieve Methoden*, 4, 1983b, 9, p. 96-112.
- Schouten, H.J.A., 'Measuring pairwise interobserver agreement, when all subjects are judged by the same observers', *Statistica Neerlandica*, 36, 1982, p. 45-61.
- Scott, W.A., 'Reliability of content analysis: the case of nominal scale coding', *Public Opinion Quarterly*, 19, 1955, p. 321-325.
- Sneath, P.H.A., & Sokal, R.R., *Numerical taxonomy*, San Francisco, Freeman, 1973.
- Varga, K., 'Achievement motivation research in Hungary', in: Halmos, P. (red.), 'Hungarian sociological studies', *The Sociological Review Monograph*, 17, University of Keele. Keele, 1972, p. 175-206.
- Vos, H.H. de, *Het meten van werkorientaties. Een vergelijking van verschillende technieken voor het meten van houdingen*, Meppel, 1980, dissertatie.
- Wiedemann, C.F. & Fenster, C.A., 'The use of chance corrected percentage of agreement to interpret the results of a discriminant analysis', *Educational and Psychological Measurement*, 38, 1978, p. 29-35.
- Woodward, J.L. en Franzen, R., 'A study on coding reliability', *Public Opinion Quarterly*, 12, 1948, p. 253-257.