12

Beoordelaarsovereenstemming

Vaak wordt bij het vaststellen van de mate waarin personen of objecten bepaalde kenmerken of eigenschappen bezitten, gebruik gemaakt van twee of meer terzake kundige beoordelaars die onafhankelijk van elkaar te werk gaan. In dergelijke gevallen nemen beoordelaars als het ware de plaats in van items of vragen in een toets of vragenlijst. Denk bijvoorbeeld aan de beoordeling van de kwaliteit van een scriptie, de beoordeling van een sportprestatie, de beoordeling van de geluidskwaliteit van stereo--apparatuur. Per beoordeelde eenheid beschikt men dan over twee of meer beoordelingen of scores. Hoewel te verwachten is dat beoordelaars niet altijd hetzelfde oordeel over een object geven, is bij grote verschillen tussen beoordelaars de bruikbaarheid van de beoordelingsprocedure twijfelachtig.

Wanneer ervaren radiologen aan de hand van röntgenfoto's de kwaadaardigheid van maagzweren beoordelen, blijkt in het algemeen dat ze lang niet altijd tot dezelfde conclusie komen (De Groot, 1966; Hofstee, 1981). Wanneer een patiënt door een arts wordt onder-zocht, is het gewenst dat diens bevindingen (diagnose, geconstateerde symptomen) niet anders luiden dan die van een andere arts die de patiënt onderzoekt. Verschillen tussen artsen impliceren dat in de praktijk sommige patiënten onnodig zullen worden geopereerd, terwijl andere patiënten een noodzakelijke, wellicht levensreddende, ingreep moeten ontberen.

In het onderwijs wordt de objectieve beoordeling van leerlingprestaties nagestreefd. Met objectief wordt bedoeld dat de uitkomst van de beoordeling slechts afhangt van de kwaliteit van de geleverde prestatie en dat ongeacht de beoordelaar hetzelfde beoordelingsresultaat wordt verkregen. Wanneer docenten echter opstellen Nederlands beoordelen, blijken voor één en hetzelfde opstel hun cijfers soms te verschillen van het cijfer 4 tot en met het cijfer 8. Dat betekent dat in examensituaties sommige leerlingen ten onrechte zakken of slagen.

Genoemde voorbeelden illustreren welke consequenties verschillen, of het gebrek aan overeenstemming tussen beoordelaars, kunnen hebben voor personen of objecten die beoordeeld worden. De voorbeelden geven tevens de relevantie aan van onderzoek

waarmee het mogelijk is (het gebrek aan) overeenstemming tussen beoordelaars, of de kwaliteit van beoordelingsprocedures te kwantificeren.

In paragraaf 12.1 van dit hoofdstuk wordt het begrip beoordelaarsovereenstemming gedefinieerd. De keuze van een maat voor beoordelaarsovereenstemming hangt af van het meetniveau van de data. In de paragrafen 12.2, 12.3 en 12.4 worden maten voor beoordelaarsovereenstemming bij data van respectievelijk nominaal, ordinaal en intervalniveau behandeld. In paragraaf 12.5 wordt een overzicht gegeven van mogelijke oorzaken voor lage beoordelaarsovereenstemming en remedies daarvoor. Tenslotte worden in paragraaf 12.6 nog een aantal andere ontwikkelingen aan de orde gesteld.

12.1 Definitie van beoordelaarsovereenstemming

Beoordelaars die oordelen geven, verrichten een beoordelingstaak. Deze taak kan opgevat worden als het classificeren van objecten. Daarmee wordt bedoeld het toewijzen van objecten aan beoordelingscategorieën op basis van een of meer -gepercipieerde- eigenschappen van die objecten. De categorieën in het eerder genoemde voorbeeld van de beoordeling van tumoren zijn bijvoorbeeld 'goedaardig', 'twijfelachtig', 'kwaadaardig'. Bij de beoordeling van prestaties van leerlingen in het onderwijs worden de categorieën gevormd door de bekende cijferschaal 1 tot en met 10. Bij beoordelingen veronderstellen we dus steeds een classificatie-schema dat een verzameling categorieën omvat. Beoordelaarsovereenstemming definiëren we als 'gelijkheid van classificatie' (Popping, 1983). De term gelijkheid in deze omschrijving is van fundamenteel belang. Daarmee wordt bedoeld dat de classificaties die door beoordelaars aan een object gegeven worden identiek zijn. We spreken van volledige overeenstemming tussen twee beoordelaars (ten aanzien van een object), als ze beiden het object toewijzen aan precies dezelfde categorie uit het classificatieschema. Deze (stringente) definitie impliceert dat alle beoordelaars beschikken over hetzelfde classificatieschema en dus niet de vrijheid hebben zelf hun beoordelingsschaal te kiezen.

12.2 Beoordelaarsovereenstemming bij data van nominaal niveau

Beoordelingsdata van nominaal niveau betreffen classificaties van personen of objecten in de zin van naamgeving of het toekennen van labels: 'katholiek', 'protestant', 'democraat', 'republikein', of 'CDA', 'VVD', 'D66'. Er moet gelden dat dergelijke categorieën in een classificatieschema wederzijds uitsluitend zijn: iemand kan dus niet

tegelijk protestant en katholiek zijn. Een ordening van de categorieën wordt niet verondersteld. Er kan niet worden gezegd dat 'protestant' meer of minder van 'iets' is dan 'katholiek'. Voor data van nominaal niveau bespreken we in deze paragraaf twee overeenstemmingsmaten: de proportie overeenstemming en de door Cohen (1960) voorgestelde coëfficiënt kappa.

Proportie overeenstemming

De proportie overeenstemming P_o is gedefinieerd als de verhouding van het aantal overeenstemmende oordelen en het totale aantal oordelen. Het percentage overeenstemming, $P_{\%}$, is gelijk aan $P_o \times 100$. De proportie overeenstemming wordt ook wel genoemd de ruwe (ongewogen) proportie overeenstemming. De proportie overeenstemming tussen twee beoordelaars, P_o, is gedefinieerd als:

$$P_o = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (12.1)$$

waarin:
X_i = 0 als de twee beoordelaars het niet eens zijn over object *i*,
X_i = 1 als de twee beoordelaars het wel eens zijn over object *i*,
n = het aantal objecten dat door de twee beoordelaars wordt beoordeeld.

De proportie overeenstemming geeft dus de proportie van de gevallen aan waarin twee beoordelaars het eens zijn over de categorisering van objecten en deze toewijzen aan dezelfde categorie. Het voordeel van deze index is dat ze eenvoudig te begrijpen is en eenvoudig berekend kan worden. Ofschoon het een van de meest populaire overeenstemmingsmaten is, heeft de proportie overeenstemming helaas ook een belangrijk nadeel. Bij beoordelingen zal meestal, naar we aannemen, het toeval een rol spelen. In welke mate dat het geval is, is onbekend. Een beoordelaar vergist zich wel eens, verliest soms de concentratie, wordt even afgeleid, neemt zijn taak niet serieus, raakt vermoeid of is soms niet consequent. Daardoor zullen niet alle classificaties correct zijn. Het is dan ook aannemelijk dat (twee) beoordelaars soms bij toeval tot eenzelfde oordeel komen. Het nadeel van de proportie overeenstemming is (Bartko & Carpenter, 1976, p. 309) dat ze geen rekening houdt met wat wel toevals-overeenstemming wordt genoemd.

Toevalsovereenstemming is de proportie overeenstemmende oordelen die we op basis van toeval mogen verwachten. We lichten dit toe met twee voorbeelden. In het eerste voorbeeld wordt aan twee beoordelaars gevraagd *n* objecten te beoordelen op een driepuntsschaal. Zij doen dat, onafhankelijk van elkaar, maar nemen hun taak volstrekt niet serieus. Elk van hun scores (categorietoewijzingen) wordt dus geheel door het toeval bepaald en heeft niets met de eigenschap van de beoordeelde objecten te maken. In tabel 12.1 hebben we de classificaties van de twee beoordelaars samengevat. De negen cellen van tabel 12.1 bevatten proporties. De proportie objecten die door de eerste beoordelaar aan categorie 1 en door de tweede beoordelaar aan categorie 2 is toegewezen (.08), staat in de gearceerde cel 1,2. De diagonaal bevat de proportie gevallen waarin identieke oordelen zijn gegeven.

Tabel 12.1
Hypothetische proporties ter illustratie van toevalsovereenstemming

		Beoordelaar 2			
	Categorie	1	2	3	Totaal
Beoordelaar 1	1	.01	.08	.01	.10
	2	.08	.64	.08	.80
	3	.01	.08	.01	.10
	Totaal	.10	.80	.10	1.00

In dit fictieve voorbeeld zien we dat zelfs bij willekeurige toewijzing van objecten, uitsluitend en alleen op basis van toeval, een hoge proportie overeenstemming kan worden verkregen. De proportie ruwe overeenstemming is hier .66, namelijk de som van de proporties op de diagonaal van de tabel. Bij het optreden van toevalsovereenstemming (Popping, 1983, p. 25, Cohen, 1960, p. 38) speelt het aantal beschikbare beoordelingscategorieën een rol, alsmede de situatie waarin beoordelingscategorieën door beoordelaars moeilijk van elkaar zijn te onderscheiden (Schouten, 1985, p. XV).

In het tweede voorbeeld wordt aan twee andere beoordelaars gevraagd *n* objecten te beoordelen op een driepuntsschaal. Zij doen dat uiterst consciëntieus en hun toewijzing van objecten aan categorieën heeft uitsluitend betrekking op de eigenschap van de beoordeelde objecten. In tabel 12.2. vatten we de gegevens samen.

Tabel 12.2
Hypothetische proporties ter illustratie van overeenstemming

		Beoordelaar 4			
	Categorie	1	2	3	Totaal
Beoordelaar 3	1	.24	.13	.03	.40
	2	.05	.20	.05	.30
	3	.01	.07	.22	.30
	Totaal	.30	.40	.30	1.00

Bekijken we de diagonaal van overeenstemmingstabel 12.2, dan stellen we vast dat ook in dit geval de proportie overeenstemming uitkomt op .66, ofschoon we toch een beduidend ander beoordelaarsgedrag veronderstellen. We moeten dan ook concluderen dat de index 'proportie overeenstemming' geen rekening houdt met toevalsovereenstemming. De proportie toevals-overeenstemming wordt bepaald op basis van de marginale proporties. Tabel 12.3 geeft de verwachte celproporties gebaseerd op de marginale proporties in tabel 12.2 bij statistische onafhankelijkheid van beoordelaars. De waarde in de gearceerde cel 1.1 met waarde .12 wordt bijvoorbeeld verkregen als het product van de rij- en kolomtotalen: .40 × .30 = .12.
We zien in tabel 12.3 dat alleen al een proportie overeenstemming van .33, de som van de diagonaalcellen, te verwachten is op basis van de marginale proporties. Dat stelt de eerder gevonden proportie overeenstemming van .66 in tabel 12.2 in een ander licht.

Tabel 12.3
Verwachte celproporties bij onafhankelijkheid van beoordelaars

		Beoordelaar 4			
	Categorie	1	2	3	Totaal
Beoordelaar 3	1	.12	.16	.12	.40
	2	.09	.12	.09	.30
	3	.09	.12	.09	.30
	Totaal	.30	.40	.30	1.00

Resumerend stellen we vast dat de proportie overeenstemming weliswaar eenvoudig te bepalen is, maar als belangrijk bezwaar heeft dat ze geen rekening houdt met toevals-overeenstemming. Cohen (1960) heeft een index voorgesteld die aan dit probleem tegemoet komt.

Coëfficiënt kappa

Coëfficiënt kappa, κ, wordt algemeen aanbevolen als maat voor het bepalen van de overeenstemming tussen twee beoordelaars. Deze overeenstemmingsindex houdt rekening met toevalsovereenstemming en is toepasbaar bij zowel dichotome als polytome data van nominaal meetniveau. Kappa kan ook gegeneraliseerd worden naar situaties met meer dan twee beoordelaars. De berekening van κ veronderstelt dat de categorieën in het classificatieschema functioneel zijn. Daarmee wordt bedoeld dat het niet is toegestaan dat er categorieën in het schema voorkomen die door een beoordelaarspaar in het geheel niet worden gebruikt. Als dat het geval is dient het classificatieschema te worden herzien.

Coëfficiënt κ wordt, net als P_o in formule (12.1), berekend op basis van een zogenaamde overeenstemmingstabel waarin de classificaties van twee beoordelaars tegen elkaar worden afgezet. Een overeenstemmingstabel (zie ook tabel 12.1 en 12.2) bevat evenveel rijen als kolommen, namelijk *c*, het aantal beschikbare categorieën in het classificatieschema. De cellen bevatten proporties. Cel P_{ij} bevat de proportie objecten die door beoordelaar 1 aan categorie *i* en door beoordelaar 2 aan categorie *j* zijn toegewezen. De diagonaal bevat de proportie gevallen waarin identieke oordelen zijn gegeven. De algemene gedaante van een overeenstemmingstabel is gegeven in tabel 12.4.

Tabel 12.4
Overeenstemmingstabel

		Beoordelaar 2					
		1	2	.	*j*	.	*c*	
Beoordelaar 1	1	P_{11}	P_{12}				P_{1c}	P_1 .
	2	P_{21}						P_2 .
	.							.
	i				P_{ij}			P_i .
	.							.
	c	P_{c1}						P_c .
		$P_{\cdot 1}$	$P_{\cdot 2}$	.	$P_{\cdot j}$	.	$P_{\cdot c}$	*n*

De verschillende symbolen in tabel 12.4 hebben de volgende betekenis:

c = het aantal beoordelingscategorieën,
n = totaal aantal beoordeelde objecten (werkstukken, personen),
i = categorie-index voor beoordelaar 1, met $i = 1, ..., c$,
j = categorie-index voor beoordelaar 2, met $j = 1, ..., c$,
P_{ij} = proportie objecten toegewezen aan categorie i en j,
$P_{i.}$ = proportie objecten toegewezen aan categorie i,
$P_{.j}$ = proportie objecten toegewezen aan categorie j.

Om κ te berekenen moet voor de overeenstemmingstabel die men wil gebruiken gelden dat $n \geq 2$ en $c \geq 2$. Er moeten dus twee of meer objecten en twee of meer categorieën zijn. De berekening van κ is niet mogelijk wanneer zowel $P_{i.}$ als $P_{.j} = 0$ (met $i = j$), in welk geval een categorie in het classificatieschema niet wordt benut. Coëfficiënt kappa is gedefinieerd als:

$$\kappa = (P_o - P_e) / (1 - P_e). \qquad (12.2)$$

In (12.2) is de geobserveerde proportie overeenstemming, P_o, gedefinieerd als:

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

Toevalsovereenstemming nulmodel is gedefinieerd als: $P_e = \sum_{i=1}^{c} P_{i.} P_{.i}$.
Coëfficiënt κ is een index voor beoordelaarsovereenstemming die, om Cohen (1960, p. 40) te citeren ..."the proportion of agreement after chance agreement is removed from consideration" weergeeft.

Keren we terug naar de overeenstemmingstabel 12.1 en we berekenen κ, dan vinden we $P_o = .66$ en $P_e = .66$, zodat $\kappa = (P_o - P_e) / (1 - P_e) = (.66 - .66)/(1 - .66) = 0 / .31 = 0$. Met andere woorden: alle waargenomen overeenstemming blijkt toevalsovereenstemming te zijn. Kijken we naar het eerder gegeven tweede voorbeeld, de serieuze beoordelaars in tabel 12.2 (en tabel 12.3) en we berekenen κ, dan vinden we $P_o = .66$ en $P_e = .33$, zodat $\kappa = (P_o - P_e) / (1 - P_e) = (.66 - .33) / (1 - .33) = .33 / .67 = .49$. De proportie overeen-stemming na correctie voor toevalsovereenstemming bedraagt dus .49. Uit de twee voorbeelden blijkt dus nog eens dat de proportie overeenstemming een onjuist beeld van de beoordelaarsovereenstemming kan geven.

De interpretatie van coëfficiënt kappa

Coëfficiënt κ is gelijk aan 1 bij perfecte overeenstemming. Een positieve waarde van κ geeft aan dat beoordelaars vaker met elkaar overeenstemmen dan op basis van toeval mag worden verwacht. Een κ van 0 geeft aan dat de mate van overeenstemming tussen beoordelaars gelijk is aan het kansniveau. Een negatieve waarde van κ geeft aan dat de beoordelaars minder vaak met elkaar overeenstemmen dan op basis van toeval kan worden verwacht, een κ van -1 wijst op een totaal gebrek aan overeenstemming tussen beoordelaars. In de literatuur wordt wel aangegeven dat een κ van .60 als een minimum moet worden beschouwd om van een acceptabele beoordelaarsovereenstemming te kunnen spreken, terwijl een κ waarde van .80 of hoger als 'goed' of 'bevredigend' wordt gekarakteriseerd (Dunn, 1989; Popping, 1983). Muskens (1980, p. 131) noemt deze grenswaarde van .80, een 'convention of the trade'. Landis en Koch (1977, p. 265) stelden het onderstaande, vaak geciteerde, overzicht op voor de interpretatie van κ.

κ			Interpretatie
<.00	<		'poor'
.00	-	.20	'slight'
.21	-	.40	'fair'
.41	-	.60	'moderate'
.61	-	.80	'substantial'
.81	-	1.00	'almost perfect'

Met betrekking tot de hoogte van coëfficiënt kappa moet opgemerkt worden dat het alleen bij gelijke marginale verdelingen in de overeenstemmingstabel mogelijk is dat kappa een maximum van 1.00 bereikt (Bartko & Carpenter, 1976, p. 314). Vandaar dat Dunn (1989, p. 38) voorstelt om bij de interpretatie de gevonden κ coëfficiënt te relateren aan de maximaal bereikbare κ, gegeven de randtotalen van de overeenstemmingstabel. Andere aspecten ten aanzien van de interpretatie van κ worden besproken door Umesh, Peterson en Sauber (1989).

Overeenstemming en associatie

In tabel 12.5 is geteld hoe twee beoordelaars honderd objecten toewijzen aan een van vier beschikbare nominale categorieën in een classificatieschema.

Tabel 12.5

Hypothetische frequenties van honderd objecten

		Beoordelaar 2				
	Categorie	1	2	3	4	Totaal
Beoordelaar 1	1	0	25	0	0	25
	2	0	0	0	25	25
	3	25	0	0	0	25
	4	0	0	25	0	25
	Totaal	25	25	25	25	100

De diagonaal in de tabel bevat alleen maar nullen, wat betekent dat het geen enkele keer voorkomt dat de twee beoordelaars een object aan dezelfde categorie toewijzen. Dit is een geval van perfecte niet-overeenstemming. Nochtans weten we dat als de eerste beoordelaar een object toewijst aan categorie 1, de tweede beoordelaar het object aan categorie 2 toewijst. Er is in dit geval sprake van perfecte samenhang of associatie. Perfecte associatie houdt in dat uit de categorie waaraan de ene beoordelaar het object toewijst, voorspeld kan worden aan welke categorie de andere beoordelaar het object toewijst. Voor één tabel kan dus gelden dat de associatie hoog is en de overeenstemming laag. Het omgekeerde geldt niet: indien er sprake is van overeenstemming geldt er ook associatie. In tabel 12.6 is er sprake van perfecte associatie, maar ook van perfecte overeenstemming.

Tabel 12.6

Hypothetische frequenties van honderd objecten

		Beoordelaar 2				
	Categorie	1	2	3	4	Totaal
Beoordelaar 1	1	25	0	0	0	25
	2	0	25	0	0	25
	3	0	0	25	0	25
	4	0	0	0	25	25
	Totaal	25	25	25	25	100

We zien in tabel 12.6 dat als we weten aan welke categorie de eerste beoordelaar een object toewijst, we ook weten aan welke categorie de tweede beoordelaar het object toewijst. We zien echter ook, dat anders dan in tabel 12.5, alle frequenties op de

diagonaal van de tabel liggen . Dat wil zeggen dat elk object door de twee beoordelaars aan dezelfde categorie (1, 2, 3 of 4) wordt toegewezen. Er is sprake van perfecte beoordelaarsovereenstemming.

Ofschoon tabel 12.5 perfecte niet-overeenstemming laat zien, wijst het voorkomen van associatie er op dat er toch een bepaalde samenhang is tussen de oordelen van de beoordelaars. Een nadeel van κ is dat alle gevallen van niet-overeenstemming gelijk worden behandeld omdat alleen naar de proporties op de diagonaal van de overeenstemmingsmatrix wordt gekeken. Daarom heeft Cohen (1968) een overeenstemmingsindex voorgesteld die aan dit bezwaar tegemoet komt. Deze index bespreken we in de volgende paragraaf.

12.3 Beoordelaarsovereenstemming bij data van ordinaal niveau

Beoordelingsdata van ordinaal meetniveau betreffen vaak beoordelingen naar de mate van aanwezig zijn van een eigenschap of kenmerk. Denk daarbij bijvoorbeeld aan Likertschalen, waarbij gegradeerde kwalificaties gegeven worden zoals 'slecht', 'matig', 'redelijk', 'voldoende', 'goed'. We spreken dan over een classificatieschema met geordende categorieën, waarbij overigens geen gelijke afstanden tussen de schaalpunten worden verondersteld. Deze ordening maakt het mogelijk rekening te houden met de mate van niet-overeenstemming. Daartoe maken we gebruik van het begrip gedeeltelijke of partiële overeenstemming. Twee beoordelaars die een object respectievelijk classificeren als 'voldoende' en 'goed' stemmen meer met elkaar overeen dan twee beoordelaars die een object beoordelen als respectievelijk 'slecht' en 'goed'.

Gewogen coëfficiënt kappa

Een maat voor beoordelaarsovereenstemming bij data van ordinaal meetniveau is de gewogen coëfficiënt kappa κ_w. Twee kenmerken van deze coëfficiënt zijn dat niet alleen gecorrigeerd wordt voor de mate van overeenstemming tussen beoordelaars die op basis van louter toeval verwacht kan worden, maar dat ook met partiële overeenstemming rekening wordt gehouden. Voor dat laatste wordt een gewichtenmatrix gebruikt. Een voorbeeld van een gewichtenmatrix staat in tabel 12.7.

Tabel 12.7

Gewichtenmatrix voor κ_w

	1	2	.	j	.	c
1	w_{11}	w_{12}				w_{1c}
2	w_{21}					
.						
i				w_{ij}		
.						
c	w_{c1}					

De symbolen in tabel 12.7 hebben de volgende betekenis:

c = het aantal beoordelingscategorieën,

i = categorie-index voor beoordelaar 1, met $i = 1, ..., c$,

j = categorie-index voor beoordelaar 2, met $j = 1, ..., c$,

w_{ij} = gewicht behorend bij toewijzingen aan categorie i en j.

De gewichten in de matrix moeten liggen tussen 0 en 1. Cellen die volledige overeenstemming representeren (gelijke classificaties) geven we het gewicht 1. Het gewicht 1 moet daarom altijd worden toegekend aan cellen die op de diagonaal van de matrix liggen, dus $w_{ii} = 1$. Het gewicht 0 wordt toegekend aan cellen die volledige niet-overeenstemming betreffen (classificaties die maximaal verschillen). Verder moet de gewichtenmatrix symmetrisch zijn ($w_{ij} = w_{ji}$) en er moet gelden $0 \leq w_{ij} \leq 1 = w_{ii}$.

Indien in de gewichtenmatrix alle cellen op de diagonaal het gewicht 1 bevatten en alle overige cellen het gewicht 0, is de gewogen coëfficiënt kappa gelijk aan κ. Coëfficiënt κ kan dan ook als een speciaal geval van κ_w opgevat worden. Beschouw nu tabel 12.8.

Tabel 12.8

Beoordeling door twee beoordelaars van werkstukken van vijf personen op een beoordelingsschaal (1 = matig; 2 = redelijk; 3 = uitstekend)

persoon	beoordelaar 1	beoordelaar 2
1	1	1
2	2	2
3	1	2
4	1	2
5	3	3

We geven nu eerst de bij deze tabel behorende overeenstemmingstabel 12.9.

Tabel 12.9

Overeenstemmingstabel van classificaties van twee beoordelaars van werkstukken van vijf personen

		Beoordelaar 2			
		1	2	3	
Beoordelaar 1	1	.20	.40	.00	.60
	2	.00	.20	.00	.20
	3	.00	.00	.20	.20
		.20	.60	.20	$n = 5$

De definitie van κ_w is: $\kappa_w = (P_o - P_e) / (1 - P_e)$ (12.3)

waarin $P_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{ij}$ de gewogen proportie overeenstemming is die we observeren

en $P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{i\cdot} P_{\cdot j}$ de gewogen proportie toevalsovereenstemming is.

De bepaling van de gewichten in de buitendiagonale cellen van de gewichtenmatrix kan op verschillende manieren gebeuren. We noemen er drie. In de eerste methode krijgen (net als de diagonale cellen) bepaalde buitendiagonale cellen op inhoudelijke

gronden het gewicht 1, de andere het gewicht 0. Dit is het geval wanneer een onderzoeker bijvoorbeeld bij nader inzien van mening is dat categorieën met verschillende labels in feite toch hetzelfde kenmerk van een object representeren. Dit is equivalent aan een hercodering van de data, waarbij categorieën worden samengevoegd. Een voorbeeld van een op deze wijze opgestelde gewichtenmatrix bij overeenstemmingstabel 12.9 geeft tabel 12.10.

Tabel 12.10
Voorbeeld van een gewichtenmatrix van κ_w

	1	2	3
1	1.00	1.00	.00
2	1.00	1.00	.00
3	.00	.00	1.00

Hier zien we dat door de gewichtentoekenning in feite de categorieën 1 en 2 worden samengenomen. De tweede methode bestaat uit het via een algoritme bepalen van zogenaamde lineaire gewichten. Dergelijke gewichten, onder andere voorgesteld door Cicchetti (1972, p. 17), worden bepaald volgens de regel:

$$w_{ij} = 1 - \left\{ |i - j| \ / \ |c - 1| \right\}.$$

Het gewicht 1 wordt toegekend aan cellen die betrekking hebben op volledige overeenstemming, waarbij dus de twee beoordelaars een object aan dezelfde categorie toewijzen. Het gewicht 0 wordt toegekend aan die cellen waarbij de (scores van) twee beoordelingen maximaal verschillen. Toepassing van deze regel op tabel overeenstemmingstabel 12.9 geeft tabel 12.11.
Het lineair gewicht in de gearceerde cel w_{12} wordt berekend als

$$w_{12} = 1 - \left\{ |1 - 2| \ / \ |3 - 1| \right\} = 1 - (1 \ / \ 2) = .50.$$

Tabel 12.11
Voorbeeld van een matrix met lineaire gewichten

	1	2	3
1	1.00	.50	.00
2	.50	1.00	.50
3	.00	.50	1.00

Bij de derde methode worden zogenaamde kwadratische gewichten (Cohen, 1968) aan de buitendiagonale cellen toegekend. Een onderzoeker vindt bijvoorbeeld dat een relatief kleine afstand tussen beoordelaars als een behoorlijke mate van overeenstemming kan worden beschouwd, maar een grotere afstand nauwelijks meer mag meetellen. Kwadratische gewichten worden bepaald volgens de regel:

$$w_{ij} = 1 - \left\{(i-j)^2 / (c-1)^2\right\}.$$

Toepassing van deze regel op overeenstemmingstabel 12.9 geeft tabel 12.12.

Tabel 12.12
Voorbeeld van een matrix met kwadratische gewichten

	1	2	3
1	1.00	.75	.00
2	.75	1.00	.75
3	.00	.75	1.00

Het kwadratisch gewicht in de gearceerde cel w_{12} wordt berekend als

$$w_{12} = 1 - \left\{(1-2)^2 / (3-1)^2\right\} = 1 - (1 / 4) = .75.$$

We geven nu een voorbeeld van de berekening van κ_w waarbij gebruik wordt gemaakt van lineaire gewichten. Tabel 12.8 bevat de ruwe data voor twee beoordelaars die van vijf personen de kwaliteit van een werkstuk beoordeelden. Elk werkstuk is aan een van $c = 3$ beoordelingscategorieën toegewezen. Tabel 12.9 is de

overeenstemmingstabel en tabel 12.11 bevat de lineaire gewichten. De proportie gewogen overeenstemming, P_o, berekenen we als:

$$P_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{ij} = w_{11}P_{11} + w_{12}P_{12} + w_{13}P_{13} + w_{21}P_{21} + w_{22}P_{22} + w_{23}P_{23} + w_{31}P_{31} + w_{32}P_{32} + w_{33}P_{33} = .20 + .20 + .00 + .00 + .20 + .00 + .00 + .00 + .20 = .80.$$

De proportie gewogen toevalsovereenstemming P_e is:

$$P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij} P_{i\cdot} P_{\cdot j} = .56.$$

De gewogen coëfficiënt kappa, κ_w, met lineaire gewichten, is gelijk aan:

$$\kappa_w = (P_o - P_e) / (1 - P_e) = (.80 - .56) / (1 - .56) = .24 / .44 = .55.$$

Merk op dat voor de data in tabel 12.9 de ongewogen coëfficiënt κ gelijk is aan .44, waarbij P_o = .60 en P_e = .28. Het is eenvoudig in te zien dat weging altijd leidt tot een waarde voor de overeenstemmingsindex die gelijk is aan of hoger is dan de ongewogen kappa. Zouden we kwadratische gewichten hebben toegepast, dan zou gewogen kappa .67 hebben bedragen, met P_o = .90 en P_e = .70.

Betrouwbaarheidsinterval voor kappa

De variantie van κ_w, $\sigma^2_{\kappa_w}$ (voor twee beoordelaars), is (Fleiss, Cohen & Everitt, 1969; Popping, 1983, 1992):

$$\frac{\sum_{i=1}^{c} \sum_{j=1}^{c} P_{ij} [(1 - P_e) w_{ij} - (1 - P_o)(w_{i\cdot} + w_{\cdot j})]^2 - (P_o P_e - 2P_e + P_o)^2}{n(1 - P_e)^4}$$

waarin $w_{i\cdot} = \sum_{j=1}^{c} w_{ij} P_{\cdot j}$ en $w_{\cdot j} = \sum_{i=1}^{c} w_{ij} P_{i\cdot}$.

Op basis van deze variantie kunnen de betrouwbaarheidsgrenzen voor kappa berekend worden. De betrouwbaarheidsgrenzen voor kappa geven aan binnen welke waarden kappa kan varieren, wanneer we het onderzoek met andere beoordelaars zouden herhalen. Deze grenzen worden bij benadering (Popping, 1989, p. 37) gegeven door

$$\left[\kappa_w \left(- z_{(1 - \frac{1}{2}\alpha)} \sigma_{\kappa_w} \right) , \quad \kappa_w \left(+ z_{(1 - \frac{1}{2}\alpha)} \sigma_{\kappa_w} \right) \right],$$

waarin $\sigma_{\kappa_w} = (\sigma^2_{\kappa_w})^{\frac{1}{2}}$ en z de standaard normale afwijking behorend bij gegeven significantie-niveau α is.

Coëfficiënt κ_w voor meer dan twee beoordelaars

Coëfficiënt κ_w is eenvoudig uit te breiden naar situaties dat er m beoordelaars zijn, met $m > 2$. In een situatie met meer dan twee beoordelaars zijn er $m(m - 1)/2$ oftewel $\binom{m}{2}$ paren beoordelaars die beschouwd kunnen worden. We kunnen dan bijvoorbeeld het gemiddelde van alle κ_w, $\overline{\kappa}_w$, berekenen van alle mogelijke paren beoordelaars. Popping (1983, p. 32) stelt echter voor te middelen bij het berekenen van P_o en P_e. Voor elk paar beoordelaars g en h worden dan $P_{o_{gh}}$ en $P_{e_{gh}}$ bepaald volgens formule (12.5). De gemiddelde gewogen kappa, $\overline{\kappa}_w$, is dan gelijk aan formule (12.3), met

$$P_o = \sum_{g=1}^{m-1} \sum_{h=g+1}^{m} P_{o_{gh}} \Big/ \binom{m}{2} \quad \text{en} \quad P_e = \sum_{g=1}^{m-1} \sum_{h=g+1}^{m} P_{e_{gh}} \Big/ \binom{m}{2}.$$

De variantie van $\overline{\kappa}_w$ voor meer dan twee beoordelaars is afgeleid door Popping (1983).

Aantal benodigde observaties

Cicchetti (1976) heeft onderzocht hoeveel observaties, in relatie met het aantal categorieën in het classificatieschema, vereist zijn om staat te kunnen maken op de berekende waarde voor kappa. Hij adviseert voor het aantal te beoordelen objecten: $n > 2c^2$, met c het aantal categorieën. Dus bij $c = 3$ beoordelingscategorieën moet het aantal observaties groter zijn dan 18 en bij $c = 7$ moet het aantal observaties groter zijn dan 98.

12.4 Beoordelaarsovereenstemming bij data van intervalniveau

Maten voor beoordelaarsovereenstemming bij data van intervalniveau zijn veelal gedefinieerd als ratio's van variantiecomponenten (zie ook hoofdstuk 3). In de literatuur (Haggard, 1958) worden dergelijke ratio's gewoonlijk aangeduid als intraklassecorrelatiecoëfficiënten. Shrout en Fleiss (1979) bespreken schattingen van intraklassecorrelatiecoëfficiënten voor drie soorten beoordelingssituaties. In deze paragraaf beperken we ons tot de meest voorkomende, namelijk de situatie waarbij een aselecte steekproef van objecten beoordeeld wordt door een aselecte steekproef van beoordelaars. Tabel 12.13 bevat de formele structuur van de datamatrix bij een dergelijk design.

Tabel 12.13
Datamatrix voor een gekruist design met twee factoren

		Beoordelaars					
		1	2	.	b	.	k	
Objecten	1	X_{11}	X_{12}				X_{1k}	$X_{1.}$
	2	X_{21}						$X_{2.}$
	.							.
	p				X_{pb}			$X_{p.}$
	.							.
	n	X_{n1}						$X_{n.}$
		$X_{.1}$	$X_{.2}$	.	$X_{.b}$	.	$X_{.k}$	$X_{..}$

In tabel 12.13 hebben de gebruikte symbolen de volgende betekenis:

k = aantal beoordelaars,
n = aantal beoordeelde personen of objecten,
p = index voor personen of objecten, met $p = 1, ..., n$,
b = index voor beoordelaars, met $b = 1, ..., k$,
X_{pb} = score voor object p van beoordelaar b,
$X_{p.}$ = somscore, over beoordelaars, voor object p,
$X_{.b}$ = somscore, over objecten, voor beoordelaar b,
$X_{..}$ = som van alle scores, over objecten en beoordelaars.

De beoordeling (score) van een persoon door een beoordelaar, X_{pb}, schrijven we als:

$$X_{pb} = \mu + (\mu_p - \mu) + (\mu_b - \mu) + (X_{pb} - \mu_p - \mu_b + \mu).$$

In dit lineaire model onderscheiden we naast het algemene gemiddelde μ, een persoonseffect, $\mu_p - \mu$, een beoordelaarseffect, $\mu_b - \mu$, en een residueel effect, $(X_{pb} - \mu_p - \mu_b + \mu)$. Elk van deze drie effecten of componenten heeft een variantie die we aanduiden met de term variantiecomponent.

Het schatten van variantiecomponenten

In hoofdstuk 3 is uiteengezet hoe de variantiecomponenten van een gekruist design met twee factoren geschat kunnen worden. In dat hoofdstuk is bij de berekening van de kwadratensommen uitgegaan van afwijkingsscores. Hier laten we zien dat we voor de berekening van kwadratensommen ook van de ruwe data kunnen uitgaan.

De totale kwadratensom, SS_{tot}, voor een gekruist design met twee factoren kan geschreven worden als:

$$SS_{tot} = SS_p + SS_b + SS_{res}$$

waarin:

$$SS_{tot} = \sum_{p=1}^{n}\sum_{b=1}^{k} X_{pb}^2 - \frac{X_{..}^2}{nk} \quad = \text{kwadratensom totaal}$$

$$SS_p = \frac{1}{k}\sum_{p=1}^{n} X_{p.}^2 - \frac{X_{..}^2}{nk} \quad = \text{kwadratensom personen}$$

$$SS_b = \frac{1}{n}\sum_{b=1}^{k} X_{.b}^2 - \frac{X_{..}^2}{nk} \quad = \text{kwadratensom beoordelaars}$$

$$SS_{res} = SS_{tot} - (SS_p + SS_b) \quad = \text{kwadratensom residu}$$

Door de kwadratensommen te delen door de vrijheidsgraden verkrijgen we de gemiddelde kwadratensommen:

$$MS_p = SS_p / (n-1) \quad = \text{gemiddelde kwadratensom personen}$$

$$MS_b = SS_b / (k-1) \quad = \text{gemiddelde kwadratensom beoordelaars}$$

$MS_{res} = SS_{res} / \{(n-1)(k-1)\}$ = gemiddelde kwadratensom residu.

De schattingen voor de variantiecomponenten zijn nu:

$\hat{\sigma}_p^2 = (MS_p - MS_{res}) / k$ = variantiecomponent personen

$\hat{\sigma}_b^2 = (MS_b - MS_{res}) / n$ = variantiecomponent beoordelaars

$\hat{\sigma}_{res}^2 = MS_{res}$ = variantiecomponent residu.

Beoordelaarsovereenstemmingscoëfficiënt

De beoordelaarsovereenstemmingscoëfficiënt, $\hat{\rho}^2$, voor k beoordelaars, is gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \left(\hat{\sigma}_b^2 + \hat{\sigma}_{res}^2\right) / k}. \qquad (12.4)$$

Wanneer de beoordelingen van verschillende beoordelaars perfect overeenstemmen, dus per beoordeelde persoon of object identiek zijn, dan zijn $\hat{\sigma}_b^2$ en $\hat{\sigma}_{res}^2$ gelijk aan nul en is de coëfficiënt gelijk aan 1. De variantiecomponent voor beoordelaars, $\hat{\sigma}_b^2$, geeft aan in welke mate beoordelaarsgemiddelden verschillen. Hoe lager de overeenstemming, des te groter de variantiecomponenten $\hat{\sigma}_b^2$ en $\hat{\sigma}_{res}^2$ zijn in verhouding tot $\hat{\sigma}_p^2$. Een relatief grote $\hat{\sigma}_b^2$ is minder bezwaarlijk dan een grote $\hat{\sigma}_{res}^2$ indien voor verschillen in gemiddelden gecorrigeerd kan worden. Bij volledig gebrek aan overeenstemming heeft de coëfficiënt de waarde nul.

In welke mate het aantal beoordelaars de mate van overeenstemming beïnvloedt, kan met (12.4) worden geschat door verschillende waarden van k, het aantal beoordelaars, in de noemer in te vullen. De coëfficiënt kan geïnterpreteerd worden als een schatting van de mate van overeenstemming tussen de gemiddelde beoordeling van k willekeurig gekozen beoordelaars en de gemiddelde beoordeling van k andere, eveneens willekeurig gekozen beoordelaars. Indien $k = 1$, dan is de coëfficiënt een schatting van de overeenstemming tussen de beoordelingen van één willekeurig gekozen beoordelaar en de beoordelingen van één andere, willekeurig gekozen beoordelaar. Indien $k = 2$, dan is de coëfficiënt een schatting van de gemiddelde overeenstemming tussen de gemiddelde beoordeling van twee beoordelaars en de gemiddelde beoordeling van twee andere, willekeurige beoordelaars. Formule (12.4) kan ook rechtstreeks in termen van gemiddelde kwadratensommen geschreven worden als:

$$\hat{\rho}^2 = \frac{MS_p - MS_{res}}{MS_p + (k-1)MS_{res} + k(MS_b - MS_{res})/n}.$$

Overeenstemming en betrouwbaarheid

In tabel 12.14 geven we twee fictieve voorbeelden van beoordelingen van werkstukken van tien leerlingen met behulp van een schoolcijferschaal.

Tabel 12.14

Hypothetische scores ter illustratie van verschillende niveaus van beoordelaarsovereenstemming en beoordelaarsbetrouwbaarheid

	Voorbeeld A			Voorbeeld B		
	Beoordelaar			Beoordelaar		
Werkstuk	1	2	3	4	5	6
1	1	3	5	1	1	1
2	1	3	5	2	2	2
3	2	4	6	3	3	3
4	2	4	6	3	3	3
5	3	5	7	4	4	4
6	3	5	7	5	5	5
7	4	6	8	6	6	6
8	4	6	8	7	7	7
9	5	7	9	8	8	8
10	5	7	9	9	9	9
$\bar{X}$ =	3.0	5.0	7.0	4.8	4.8	4.8
s_x =	1.5	1.5	1.5	2.7	2.7	2.7

In voorbeeld A zien we dat de drie beoordelaars steeds elk werkstuk of object een andere score geven. Van overeenstemming is dus geen sprake. We zien echter ook dat in de data een bepaald patroon zit. Beoordelaar 2 geeft steeds twee scorepunten meer dan beoordelaar 1, en beoordelaar 3 geeft steeds twee scorepunten meer dan beoordelaar 2. Het verschijnsel dat per object de scores, op een constante na, aan elkaar gelijk zijn, wordt additieve bias genoemd. De spreiding van de scores is voor elke beoordelaar gelijk. De scores van de drie beoordelaars correleren perfect met elkaar, dat wil zeggen dat elke beoordelaar tot dezelfde rangordening van werkstukken komt. In voorbeeld A is sprake van wat we perfecte beoordelaarsbetrouwbaarheid noemen. Beoordelaarsbetrouwbaarheid wordt gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{res}^2 / k}. \qquad (12.5)$$

Formule (12.5) verschilt van formule (12.4) door het ontbreken van $\hat{\sigma}_b^2$, de variantiecomponent beoordelaars. Merk op dat (12.5) gelijk is aan de definitie van Cronbachs alpha (zie hoofdstuk 3).

In voorbeeld B zien we dat de drie beoordelaars steeds elk werkstuk dezelfde, identieke, score toekennen. De gemiddelde scores van de beoordelaars en ook de spreidingen zijn gelijk. Er is hier sprake van perfecte beoordelaarsovereenstemming. We zien ook dat de scores van de drie beoordelaars perfect correleren, dus perfect betrouwbaar zijn. De twee voorbeelden laten zien dat een hoge beoordelaarsbetrouwbaarheid een noodzakelijke, maar geen voldoende voorwaarde is voor een hoge beoordelaarsovereenstemming.

Samenvattingsopdracht Nederlands

Sanders, Hendrix en Luijten (1984) trokken in het kader van hun onderzoek naar het functioneren van globale en analytische beoordelingsschema's een aselecte steekproef van dertig leerlingen die bij het centraal schriftelijk eindexamen voor het vak Nederlands een samenvattingsopdracht hadden gemaakt. Een samenvattingsopdracht houdt in dat van een langere betogende tekst een sterk verkorte, maar adequate, samenvatting moet worden gemaakt van maximaal 500 woorden. Globale beoordelingsschema's omvatten niet meer dan enkele beknopte algemene richtlijnen voor de beoordelaars. In dit geval bijvoorbeeld onder andere de instructie dat beoordeeld moet worden of de samenvatting representatief is voor de oorspronkelijke tekst en gevolgd kan worden door een lezer die de oorspronkelijke tekst niet kent. Daarbij dient de beoordelaar zijn waardering rechtstreeks uit te drukken in een cijfer. Een analytische beoordelingsschema daarentegen geeft veel meer gedetailleerde aanwijzingen en vereist dat de beoordelaar per te beoordelen aspect, zoals tekststructuur, tekstlengte, inhoud en formulering een afzonderlijke beoordelingsscore toekent. Vervolgens worden de scores op de aspecten gewogen naar hun relatieve belang en daarna samengevat in een cijfer. De dertig samenvattingen werden door acht beoordelaars onafhankelijk van elkaar beoordeeld. Tabel 12.15 bevat de resultaten van de globale beoordeling van de acht beoordelaars (B1 - B8). We zien in tabel 12.15 dat het nogal wat uitmaakt door welke beoordelaar een leerling wordt beoordeeld. Leerling 3 krijgt van beoordelaar 3 het cijfer 2.0 en van beoordelaar 8 het cijfer 6.0. Over het geheel genomen oordelen

beoordelaars 1 en 5 wat milder, terwijl beoordelaar 6 en 7 als strenge beoordelaars gekenmerkt kunnen worden. Tabel 12.16 geeft de resultaten van de variantie-analyse voor de data in tabel 12.15.

Tabel 12.15

De globale beoordeling van dertig samenvattingen door acht beoordelaars

Leerling	B1	B2	B3	B4	B5	B6	B7	B8	Som
1	6.0	6.0	8.0	6.5	9.0	6.0	7.0	7.0	55.5
2	6.5	6.0	7.0	6.0	6.5	4.0	7.0	7.0	50.0
3	4.0	5.5	2.0	5.0	3.0	4.0	4.0	6.0	33.5
4	7.5	5.0	6.0	5.0	8.5	5.0	7.0	6.0	50.0
5	6.5	4.5	4.5	4.0	6.5	4.0	4.0	6.0	40.0
6	6.0	6.0	7.0	5.5	7.5	5.0	5.0	7.0	49.0
7	7.0	5.0	3.8	5.0	7.0	4.0	6.0	7.0	44.8
8	7.0	7.5	7.0	7.0	7.0	4.0	6.0	8.0	53.5
9	7.0	6.0	6.8	6.0	7.0	5.0	6.0	7.0	50.8
10	6.5	5.0	6.8	5.5	6.5	6.0	7.0	8.0	51.3
11	8.5	7.5	7.0	8.0	10.0	7.0	5.0	9.0	62.0
12	8.0	6.0	7.5	5.5	7.5	6.0	3.0	7.0	50.5
13	7.5	6.0	6.5	6.0	7.5	7.0	6.0	6.0	52.5
14	6.0	6.0	7.0	5.5	5.0	6.0	5.0	6.0	46.5
15	8.0	6.0	6.5	6.0	6.5	6.0	3.0	6.0	48.5
16	6.5	7.0	6.5	6.5	7.0	5.0	3.0	5.0	46.5
17	9.0	5.0	7.0	5.5	7.5	4.0	7.0	7.0	52.0
18	7.5	6.0	8.0	6.5	6.5	5.0	6.0	5.0	50.5
19	7.0	5.0	6.0	5.0	8.0	6.0	5.0	6.0	48.0
20	4.0	6.5	4.0	6.0	4.5	5.0	3.0	4.0	37.0
21	4.0	6.0	3.0	6.0	4.0	5.0	4.0	4.0	36.0
22	6.0	6.0	7.0	5.5	7.5	7.0	8.0	5.0	52.0
23	4.0	4.0	5.0	4.0	4.0	5.0	7.0	6.0	39.0
24	6.5	6.0	7.0	6.5	7.5	6.0	8.0	6.0	53.5
25	7.5	6.0	8.0	6.0	5.0	5.0	6.0	4.0	47.5
26	8.0	7.5	7.5	7.0	7.0	6.0	7.0	6.0	56.0
27	5.0	4.0	4.5	3.0	6.0	3.0	3.0	5.0	33.5
28	3.0	5.0	1.0	5.0	3.0	3.0	5.0	3.0	28.0
29	5.0	4.5	6.0	4.0	5.0	4.0	6.0	5.0	39.5
30	4.0	5.5	4.0	5.0	4.0	3.0	5.0	4.0	34.5
Som	189	172	177.9	168	191.5	151	154	178	1391.4

In tabel 12.16 zien we dat de residuele component de grootste variantiecomponent is. De variantiecomponent beoordelaars daarentegen is relatief gering.

Tabel 12.16

Resultaten van de variantie-analyse voor de gegevens van de globale beoordeling van dertig werkstukken door acht beoordelaars

Effecten	Vrijheids-graden	Kwadraten-sommen	Gemiddelde kwadratensommen	Schattingen van variantiecomponenten
Personen (*p*)	29	236.31	8.15	$\hat{\sigma}^2_p$ = .876 (40%)
Beoordelaars (*b*)	7	41.05	5.86	$\hat{\sigma}^2_b$ = .157 (7%)
Residu (*res*)	203	231.92	1.14	$\hat{\sigma}^2_{res}$ = 1.143 (53%)

De beoordelaarsovereenstemmingscoëfficiënt voor $k = 8$ beoordelaars is gelijk aan:

$$\hat{\rho}^2 = \frac{.876}{.876 + (.157 + 1.143)/8} = .84 .$$

Het doel van het gebruik van beoordelingsschema's is het realiseren van een objectieve beoordeling. Dat wil zeggen dat we ernaar streven een beoordelingsschema te maken dat een zo hoog mogelijke beoordelaarsovereenstemming oplevert bij zo weinig mogelijk beoordelaars. Het zou ideaal zijn om in de beoordelingsprocedure slechts één beoordelaar in te hoeven inschakelen. In de praktijk zijn acht beoordelaars overigens meestal niet beschikbaar of betaalbaar. De geschatte overeenstemming voor het geval dat de samen-vattingen zouden worden beoordeeld door één beoordelaar is:

$$\hat{\rho}^2 = \frac{.876}{.876 + (.157 + 1.143)/1} = .40.$$

Een overeenstemmingscoëfficiënt van .40 betekent dat indien de werkstukken door één willekeurig gekozen beoordelaar beoordeeld worden, en deze beoordelingsscores zouden vergeleken worden met de scores van één andere willekeurige beoordelaar, we grote scoreverschillen zullen zien. In tabel 12.17 worden schattingen gegeven voor de overeenstemming bij gebruik van diverse aantallen beoordelaars.

In het genoemde onderzoek (Sanders et al., 1984) bleek dat met een analytisch beoordelingsschema een hogere beoordelaarsovereenstemming kon worden bereikt dan met een globaal beoordelingsschema. Bij een analytische, onafhankelijke beoordeling van samen-vattingen door twee beoordelaars kon dezelfde overeenstemming worden bereikt als met een globale beoordeling door drie onafhankelijke beoordelaars. Het behoeft geen betoog dat een beoordelingsprocedure waarin bij gelijkblijvende kwaliteit

van de beoordeling met minder beoordelaars kan worden volstaan, uit logistiek en kostenoogpunt de voorkeur verdient.

Betrouwbaarheidsinterval voor de overeenstemmingscoëfficiënt

De overeenstemmingscoëfficiënt $\hat{\rho}^2$ die we berekenen is een schatting. Bij replicaties van het onderzoek met andere steekproeven van kandidaten en beoordelaars verwachten we niet dezelfde resultaten te vinden. Het is daarom van belang het betrouwbaarheidsinterval voor de overeenstemmingscoëfficiënt ρ^2 te berekenen.

De methode voor het bepalen van een dergelijk betrouwbaarheidsinterval voor de overeenstemmingscoëfficiënt is ontleend aan Fleiss en Shrout (1978, 1979). Het betrouwbaarheidsinterval kan als volgt benaderd worden. Het aantal vrijheidsgraden, v, is gelijk aan:

$$v = \frac{(k-1)(n-1)\left\{k\hat{\rho}^2 F_b + n\left[1+(k-1)\hat{\rho}^2\right] - k\hat{\rho}^2\right\}}{(n-1)k^2\hat{\rho}^{2^2}F_b{}^2 + \left\{n\left[1+(k-1)\hat{\rho}^2\right] - k\hat{\rho}^2\right\}}.$$

In bovenstaande formule is $F_b = MS_b / MS_{res}$. Als we nu uit de *F*-verdeling de waarden definiëren $F^* = F_{1-½\alpha}[(n-1), v]$ en $F_* = F_{1-½\alpha}[v, (n-1)]$, dan zijn de grenzen van het $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval voor ρ^2:

$$\left(\frac{n(MS_p - F^* MS_{res})}{F_*[kMS_b + (kn-k-n)MS_{res}]nMS_p} \quad , \quad \frac{n(F_* MS_p - MS_{res})}{kMS_b + (kn-k-n)MS_{res} + nF_* MS_p}\right).$$

Het minimum aantal beoordelaars

De ondergrens van het betrouwbaarheidsinterval van ρ^2 is richtinggevend voor het antwoord op de vraag hoeveel beoordelaars minimaal nodig zullen zijn om in vervolgsituaties, dus bij hernieuwd beoordelen (andere kandidaten, andere beoordelaars), een bepaalde zekerheid te hebben over de te verwachten beoordelaarsovereenstemming. We zullen dat hier aan de hand van het voorbeeld van de samenvattingsopdracht Nederlands toelichten. De beoordelaarsovereenstemming voor acht beoordelaars bedroeg .84, terwijl de grenzen voor het 90% betrouwbaarheidsinterval bij benadering .76 en .91 zijn. Stel nu dat een onderzoeker aanbevelingen wil doen voor toepassing in de praktijk van de onderzochte beoordelings-

procedure, maar bijvoorbeeld, mede gelet op het kostenaspect, tevreden zou zijn met een beoordelaarsovereenstemming van .60. De beoordelaarsovereenstemmings-coëfficiënt en de daarbij geschatte betrouwbaarheidsintervallen bij verschillende aantallen beoordelaars staan in tabel 12.17. Het betreft hier opnieuw de gegevens voor de globale beoordeling van dertig samenvattingen door acht beoordelaars.

Tabel 12.17

Schattingen van de beoordelaarsovereenstemming bij diverse aantallen beoordelaars en de grenzen voor een 90% betrouwbaarheidsinterval

Aantal beoordelaars	Beoordelaars-overeenstemming	Intervalgrenzen 90% betrouwbaarheidsinterval
1	.40	.29 - .55
2	.57	.44 - .71
3	.67	.55 - .78
4	.73	.62 - .83
5	.77	.67 - .86
6	.80	.71 - .88
7	.83	.74 - .89
8	.84	.76 - .91

Inspectie van tabel 12.17 leert dat bij vier beoordelaars het interval tussen .62 en .83 ligt.
Op grond hiervan kan de conclusie worden getrokken dat voor de beoordeling van een nieuwe reeks objecten kan worden volstaan met een beoordeling door vier beoordelaars.

12.5 Lage beoordelaarsovereenstemming: oorzaken en remedies

Oorzaken

Er zijn diverse factoren denkbaar die de beoordelaarsovereenstemming nadelig beïnvloeden. Saal, Downey en Lahey (1980) geven een overzicht en merken op dat er weinig overeenstemming schijnt te bestaan over de conceptuele definities met betrekking tot de criteria voor de kwaliteit van beoordelingen en over operationele definities voor die criteria. We kunnen een onderscheid maken tussen niet-systematische en systematische invloeden. Niet-systematisch noemen we toevallige en fluctuerende invloeden op de beoordelaar en diens beoordeling. We kunnen hierbij

denken aan vermoeidheid, schrijffouten, telfouten, onoplettendheid, verstoringen van de beoordeling door lawaai en temperatuur. Systematische invloeden maken dat de beoordelingen van een beoordelaar op een systematische manier afwijken van de beoordelingen die andere beoordelaars geven.

Een bekende systematische afwijking is 'restriction of range'. Hiervan is sprake wanneer sommige beoordelaars niet alle beschikbare categorieën in een classificatieschema benutten. Twee bekende vormen hiervan zijn mildheid en centrale tendentie. Van mildheid is sprake wanneer beoordelaars de neiging hebben relatief lage of juist relatief hoge scores te geven. Zo geven sommige docenten nooit cijfers hoger dan 8 en anderen nooit cijfers lager dan 4, ongeacht het bereik van de schoolcijferschaal of de prestaties van hun leerlingen. Saal et al. (1980) geven drie operationele definities voor dit effect. Sommige beoordelaars neigen ertoe geen expliciete uitspraken te willen doen. Ze vermijden extreem geformuleerde categorieën en zitten met hun beoordelingen steeds rond het midden van de beoordelingsschaal. Dit verschijnsel wordt wel centrale tendentie genoemd.

We spreken van een halo-effect wanneer beoordelaars hun oordeel mede laten afhangen van voor de meting niet terzake doende kenmerken van degene die beoordeeld wordt of van diens product, zoals uiterlijk, kleding of de netheid van het handschrift. Zo valt de beoordeling van een prestatie of werkstuk van een vriendelijk en beleefd persoon soms hoger uit dan de beoordeling van een prestatie van een persoon die in dit opzicht afwijkt van wat de beoordelaar als normaal beschouwt. Saal et al. (1980) beschrijven het halo-effect als het onvermogen van een beoordelaar om onderscheid te maken tussen verschillende aspecten van het gedrag van de persoon die beoordeeld wordt. Ze presenteren daarbij overigens vier verschillende operationele definities.

De neiging van een beoordelaar om zich in de strengheid van zijn beoordelingen aan te passen aan het gemiddelde niveau van de te beoordelen objecten staat bekend als normverschuiving. Hoe goed of hoe slecht een schoolklas als geheel ook is voor een bepaald vak, vaak zien we dat de percentages onvoldoendes bij elke klas voor een vak gelijk zijn.

Van een sequentie-effect spreken we wanneer de beoordeling die de beoordelaar aan een object geeft mede tot stand komt op basis van de nawerking van een beoordeling die net tevoren is gegeven. De middelmatige prestatie van een leerling die wordt beoordeeld net nadat een of meer zeer slecht presterende leerlingen zijn beoordeeld, wordt dan hoger gescoord dan in het omgekeerde geval, wanneer de beoordeling van een middelmatige leerling zou volgen op de beoordeling van een of meer excellente leerlingen.

Als laatste noemen we het signifisch effect. Hiervan is sprake wanneer beoordelaars de beoordelingstaak verschillend opvatten, omdat ze de nadruk leggen op verschillende aspecten. Bij de beoordeling van het opstel zien we bijvoorbeeld dat sommige docenten meer op stijl letten, anderen op inhoud, weer anderen op structuur, terwijl de ene docent spel- en schrijffouten in de beoordeling betrekt en de andere docent weer niet.

Remedies

Constateert men een te lage beoordelaarsovereenstemming, dan zijn er verschillende manieren om er voor te zorgen dat bij herhaling van de beoordelingsprocedure betere resultaten te verwachten zijn. Bepaalde maatregelen zijn eveneens mogelijk indien herhaling van de beoordelingsprocedure niet mogelijk is. Dit laatste betreft dan met name correcties op basis van aanwijsbare systematische fouten, zoals mildheid.

Dat het inzetten van meer beoordelaars de beoordelaarsovereenstemming kan verhogen is in het voorgaande al uitvoerig besproken. Merk echter op dat ook hier de wet van de verminderende meeropbrengst van toepassing is: de winst die elke toegevoegde beoordelaar oplevert in termen van verbetering van de overeenstemming begint op een gegeven ogenblik af te nemen, meestal na twee of drie beoordelaars.

Een duidelijke verbetering van de beoordelaarsovereenstemming kan worden verwacht wanneer beoordelaars worden getraind voor hun taak, bijvoorbeeld door met hen enkele proefbeoordelingen te doen en deze te bespreken. Van de proefobjecten moet bij voorkeur het resultaat bekend zijn van een standaardbeoordeling, zodat de beoordelaars hun eigen beoordelingsscores met deze standaard kunnen vergelijken.

Men dient er voor te zorgen dat beoordelaars werkelijk onafhankelijk van elkaar werken. Overleg tussen beoordelaars gedurende de uitvoering van de beoordelingstaak draagt het risico in zich dat oneigenlijke factoren (dominantie, senioriteit, status, argumentatievermogen) het overleg en daarmee de meting beïnvloeden.

Belangrijk is ook een merkbare controle op het werk van de beoordelaars. Indien beoordelaars weten dat hun werk wordt gecontroleerd, zullen ze zich minder afwijkingen van het beoordelingsschema en de bijbehorende instructies veroorloven. In veel beoordelingssituaties komt het voor dat op een of andere wijze de beoordelaars belang hebben bij de uitslag van de beoordeling.

Beoordelaarsovereenstemming is ook afhankelijk van de kwaliteit van beoordelaars-instructies. Gezorgd dient te worden voor duidelijke en hanteerbare beoordelaarsinstructies die, bijvoorbeeld bij een beoordelaarstraining, met de beoordelaars besproken worden. Beoordelaarsinstructies hebben bijvoorbeeld betrekking op de volgorde waarin objecten worden beoordeeld, de inrichting van de beoordelingssituatie (plaats, licht, geluid), op zaken zoals 'geen aantekeningen maken op schriftelijke werkstukken' om een mogelijke tweede beoordelaar niet te beïnvloeden. Zorg daarnaast voor een helder en functioneel classificatie-schema, zodanig dat alle beoordelaars op dezelfde wijze begrijpen wat de erin voorkomende categorieën betekenen. Beperk het aantal categorieën tot maximaal zeven (James et al., 1984; Cicchetti, 1976). Belangrijk is een duidelijk scoringsvoorschrift, dat wil zeggen een overzicht van het aantal scorepunten dat gegeven dient te worden aan bijvoorbeeld een goed, een minder goed en een fout antwoord. Geef bij globale of holistische beoordelingen een overzicht waarin wordt aangegeven op welke beoordelingsaspecten gelet moet worden. Gebruik waar mogelijk analytische beoordelingsschema's. Overweeg om beoordelaars die extreem afwijkende scores te zien geven te verwijderen uit de groep beoordelaars die bij de beoordeling wordt betrokken. Is van een beoordelaar systematisch afwijkend beoordelaarsgedrag bekend, met name mildheid of strengheid, overweeg dan aanpassing van diens scores.

12.6 Tot besluit

In de beoordelingssituaties die we in dit hoofdstuk beschreven hebben, had overeenstemming altijd betrekking op overeenstemming tussen beoordelaars. Overeenstemming tussen beoor delaars wordt in de literatuur vaak aangeduid als interbeoordelaarsovereenstemming. In beoordelingssituaties waarbij één beoordelaar een reeks personen of objecten op twee verschillende tijdstippen beoordeelt, kunnen we de overeenstemming tussen de scores op de twee tijdstippen uitrekenen. In dat geval spreken we over intrabeoordelaarsovereenstemming. Wanneer er sprake is van beoordelingssituaties waarbij de overeenstemming berekend wordt tussen beoordelaars en een standaard, spreken we van accuraatheid (Suen & Ary, 1989). Deze term is ontleend aan onderzoek dat in de exacte disciplines plaatsvindt en waarbij 'echte' standaarden worden gebruikt. Zo kan de overeenstemming berekend worden tussen metingen met verschillende duimstokken ('beoordelaars') die in de handel zijn en de 'echte' meetlat of standaard. Een hoge overeenstemming tussen een bepaalde duimstok en de standaard betekent dat die duimstok valide is voor het meten van lengte. In de

sociale wetenschappen is het soms mogelijk om voor bepaalde beoordelingssituaties standaards te gebruiken, bijvoorbeeld de oordelen van enkele deskundige beoordelaars aan wiens oordeel niet getwijfeld kan worden. Het gebruik van een standaard heeft als voordeel dat beoordelingssituaties vermeden worden waarbij we een hoge beoordelaarsovereenstemming vinden terwijl de groep beoordelaars collectief verkeerd beoordeeld heeft.

De bespreking van de overeenstemmingscoëfficiënten bij data van intervalniveau beperkte zich in de vorige paragraaf tot een design met twee factoren. In paragraaf 3.13 van hoofdstuk 3 is een gekruist design met drie factoren, in de generaliseerbaarheidstheorie een design met twee facetten genoemd, besproken. Daar zagen we dat in een gekruist design met drie factoren behalve de score X_{pvb}, de score die persoon *p* voor het antwoord op vraag *v* van beoordelaar *b* ontvangen heeft, zes gemiddelde scores onderscheiden worden. Twee voorbeelden zijn de gemiddelde score van vraag *v* (gemiddeld over alle personen en alle beoordelaars) en de gemiddelde score van beoordelaar *b* (gemiddeld over alle personen en alle beoordelaars). Overeenstemmingscoëfficiënten voor designs met drie factoren zijn afgeleid door Maxwell en Pilliner (1968). Hun afleiding is gebaseerd op het concept 'replicatie van het experiment'. Dit concept gebruikt ook Mellenbergh (1977) bij zijn afleiding van wat hij replicatiecoëfficiënten noemt. Een replicatiecoëfficiënt is gedefinieerd als de correlatie tussen bijvoorbeeld de gemiddelde score van beoordelaar(s) *b* bij een beoordelingsprocedure of 'experiment' en de gemiddelde score van beoordelaar(s) *b* bij een herhaling of replicatie van de beoordelingsprocedure. Een replicatiecoëfficiënt kan geschreven worden als een ratio van variantiecomponenten. Hoe variantiecomponenten geschat kunnen worden, is uitgebreid in hoofdstuk 3 beschreven. Voor een gekruist design met drie factoren kunnen in totaal 19 replicatiecoëfficiënten geschat worden. Voor details verwijzen we naar het artikel van Mellenbergh (1977, p. 380).

In de praktijk komt het regelmatig voor dat niet alle beoordelaars alle personen (kunnen) beoordelen. Behalve het weglaten van de objecten met ontbrekende scores, bespreekt Popping (1983, p. 46) nog andere methoden om in zulke gevallen kappa te berekenen. Voor data van intervalniveau hebben Houston, Raymond en Svec (1991) een drietal methoden ontwikkeld voor het schatten van beoordelaarseffecten in het geval dat beoordelingen ontbreken. Hierdoor is het toch mogelijk te corrigeren voor verschillen in strengheid van beoordelaars. De methoden zijn verwant aan de methoden die in hoofdstuk 7 besproken zijn. Van belang is op te merken dat statistische pakketten (bijvoorbeeld Dixon, 1992) tegenwoordig programma's bevatten waarmee variantiecomponenten van incomplete gegevensverzamelingen geschat kunnen worden.

In dit hoofdstuk hebben we ons beperkt tot overeenstemmingscoëfficiënten die hun bruikbaarheid bewezen hebben. Daarnaast zijn er de laatste jaren nog vele andere overeenstemmingscoëfficiënten voorgesteld. Zegers (1991) bespreekt de eigenschappen van zogenaamde associatiecoëfficiënten. Uebersax (1991) laat zien dat het ook mogelijk is beoordelaarsovereenstemming te modelleren en te berekenen met behulp van latente klassen-modelllen, loglineaire modellen, itemresponsmodellen, correspondentie- en homogeniteits-analyse.