

DE SUPERIORITEIT VAN DE SYNTHETISCHE BEOORDELAAR

door A. W. BURGER, klinisch psycholoog

(uit de Psychiatrische Inrichting 'Het Groot Graffel' te Warnsveld,
Geneesheer-Directeur J. Westerhuis)

INLEIDING

De Duitse sociaal-psycholoog HOFSTÄTTER (1957) beschrijft het volgende experiment. Aan zeven proefpersonen worden tien zwarte figuren op een witte achtergrond aangeboden. Deze figuren verschillen van elkaar in vorm en oppervlakte, maar de oppervlakteverschillen zijn slechts gering. De proefpersonen krijgen de opdracht de figuren te rangschikken naar oppervlakte, van klein naar groot. Zij doen dit onafhankelijk van elkaar. De mate, waarin zij de werkelijke rangorde weten te benaderen, wordt uitgedrukt in een correlatie-coëfficiënt.

De prestaties blijken zeer uiteen te lopen. De coëfficiënten variëren van 0,03 tot 0,67 met een gemiddelde van 0,36. Vervolgens wordt een synthetische groep geformeerd door het combineren van de individuele uitkomsten. Dit geschiedt op de volgende wijze. Bij de zeven rangordebepalingen heeft elke figuur zeven maal een rangnummer gekregen. Per figuur worden nu de rangnummers opgeteld, zodat men tien totalen krijgt. De rangorde van deze totalen wordt weer vergeleken met de werkelijke rangorde en men krijgt thans een correlatie-coëfficiënt van 0,79!

Met behulp van een simpele berekening wordt dus een synthetische groep geconstrueerd, die meer presteert dan de enkelingen waaruit hij is opgebouwd. Het verschijnsel, dat hier zo elegant door Hofstätter in een rijtje cijfers zichtbaar wordt gemaakt, is in de psychometrie reeds lang bekend en wordt in de psychometrische praktijk dagelijks benut.

Wat is nu de waarde van dit curieuse verschijnsel voor de psychiatrie? In hoeverre is ook bij het beoordelen van patiënten de synthetische groep superieur aan de enkeling? Het doel van mijn voordracht is, een bijdrage te leveren tot de beantwoording van deze vraag aan de hand van gegevens, die met beoordelingsschalen werden verkregen.

Alvorens hiermee te beginnen, wil ik eerst enkele termen introduceren, die mij in staat moeten stellen ingewikkelde zinsconstructies te vermijden. In plaats van de term „synthetische groep” van Hofstätter zou ik voortaan de term „synthetische beoordeelaar” willen gebruiken, omdat hiermee tevens de functie van de groep wordt aangeduid. Verder zal ik spreken van tweelingen, drielingen, vier-

lingen en meerlingen, waarmee ik synthetische groepen bedoel, die respectievelijk uit twee, drie, vier of meer individuele beoordelaars zijn samengesteld.

INTERBEOORDELAAR-BETROUWBAARHEID ALS MAATSTAF

Wie benadert in de psychiatrie de werkelijkheid dichter, de synthetische of de individuele beoordelaar? Bij de beantwoording van deze vraag stuit men op een moeilijkheid, die zich in het experiment van Hofst tter niet voordoet. Men weet namelijk in de psychiatrie niet, wat de werkelijkheid is. Niemand kan achter de schermen kijken en verklappen hoe de werkelijke rangschikking van een aantal pati nten volgens een bepaald gedragsaspect is.

Men kan het werkelijkheidsgehalte van een beoordeling slechts afleiden uit de mate, waarin zij overeenstemt met een beoordeling, die door iemand anders over dezelfde pati nten wordt geproduceerd. Is er een vrijwel volledige overeenstemming, dan heeft men blijkbaar te doen met een onmiskenbaar, ondubbelzinnig stuk werkelijkheid. De mate van overeenstemming wordt uitgedrukt in een interbeoordelaar-betrouwbaarheidsco ffici nt. Deze geeft de correlatie aan tussen twee series beoordelingen, die beide betrekking hebben op dezelfde reeks gevallen waarin een bepaald verschijnsel zich in verschillende gradaties voordoet, maar welke series afkomstig zijn van verschillende beoordelaars.

Hoe hoger de co ffici nt, hoe meer de beoordelaars overeenstemmen en hoe dichter zij tezelfdertijd de werkelijkheid benaderen. Uit de co ffici nt is dus ook de kwaliteit van de beoordelaars af te leiden. Een slechte beoordelaar blijft verder van de realiteit verwijderd en zal dus minder overeenstemming vertonen met andere beoordelaars dan de goede beoordelaar.

Dit maakt het dus mogelijk de synthetische en de individuele beoordelaar met elkaar te vergelijken. Het feit, dat deze vergelijking moet geschieden met behulp van betrouwbaarheidsco ffici nten, onderstreept het grote praktische belang van deze vergelijking. Immers, de interbeoordelaar-betrouwbaarheid neemt een sleutelpositie in. De wetenschap zoekt primair naar verschijnselen, waarover een grote mate van eenstemmigheid mogelijk is. Alleen deze verschijnselen kunnen een bruikbaar uitgangspunt vormen voor een voortgezet onderzoek, b.v. naar het verband van deze verschijnselen met andere.

GEGEVENS UIT DE LITERATUUR

JENSEN en MORRIS (1960) beschrijven een onderzoek op vijf verschillende verpleegafdelingen, waar zij zes tot acht beoordelaars de pati nten hebben laten rangschikken op een apathie-agitatie conti-

nuum. De correlatie tussen de individuele beoordelaars is per afdeling gemiddeld als volgt: 0,78; 0,70; 0,56; 0,48; 0,46. Vervolgens hebben dezelfde beoordelaars dezelfde patiënten beoordeeld met een vijfpunts-apathieschaal. De gemiddelden per afdeling zijn nu: 0,52; 0,58; 0,71; 0,65; 0,44.

De auteurs schatten tenslotte (met behulp van de Spearman-Brown formule), wat de constructie van synthetische beoordelaars zou opleveren. Bij een combinatie per afdeling van de uitkomsten van beide onderzoeken en van alle beoordelaars zouden de volgende enorme cijfers ontstaan: 0,98; 0,97; 0,95; 0,94; 0,91.

Jensen en Morris vermelden nog, dat een tweeling, bestaande uit twee psychologen of twee psychiaters, evenveel presteert als een zevenling van verplegenden. Mijns inziens geldt dit alleen voor een diagnostische schaal; bij een aanpassingsschaal zal de superioriteit van de academici zeker niet zo groot zijn.

BEYAERT (1966) heeft de Nederlandse psychiatrie verrast met een aantal indrukwekkend hoge interbeoordelaar-betrouwbaarheids-cijfers, welke hij heeft verkregen met de PUK-schalen. De meeste liggen in de buurt van de 0,90. Ongetwijfeld is deze prestatie mede te danken aan het feit, dat de auteur uitsluitend synthetische beoordelaars heeft gebruikt: een zesling, een vijfling, vier vierlingen en een achtling.

EIGEN GEGEVENS

Vier beoordelaars hebben onafhankelijk van elkaar dezelfde vijf en twintig patiënten beoordeeld met de Warnsveldse Aanpassingsschaal voor Verpleegafdelingen (WAV), welke op het ogenblik aan „Het Groot Graffel” wordt ontwikkeld door A. W. Burger, R. Hofman, N. J. Peters en F. J. Wolters. Het is een schaal, die met de reeds eerder ontwikkelde Graffelschaal (BURGER, DE WAARD, WILDE 1966) vier subschalen gemeen heeft en in constructief opzicht gelijk is aan deze schaal. Zij bestaat uit acht vijftraps-subschalen en levert acht subscores, die minimaal 1 en maximaal 5 kunnen bedragen, en een eindscore, welke de som is van de subscores.

De beoordeelde patiënten zijn allen mannen. Hun leeftijden variëren van 21 tot 68 met een gemiddelde van ongeveer 40. Een kwart van de patiënten vertoeft reeds jarenlang in de inrichting, de rest behoort tot de doorstromende bevolking. Meer dan de helft is schizofreen.

Drie van de beoordelaars zijn tweede- en derdejaars leerling-verplegers, de vierde is gediplomeerd verpleger. Bij de aanvang van het onderhavige onderzoek waren zij reeds enigmate vertrouwd met de WAV doordat zij al eerder enige tientallen WAV-beoordelingen hadden gemaakt.

Ons onderzoek heeft een groot aantal interbeoordelaar-betrouwbaarheidscoëfficiënten (Pearson-r) opgeleverd, waarvan ik alleen die zal weergeven, welke betrekking hebben op de eindscores en op de scores van de subschalen A, B, C en D. Dit zijn de vier nieuwe subschalen, die de WAV niet gemeen heeft met de Graffelschaal. De cijfers vindt men in tabel 1.

TABEL 1. De prestaties van individuele en synthetische beoordelaars uitgedrukt in interbeoordelaar-betrouwbaarheidscoëfficiënten verkregen met vier subschalen van de WAV (N = 25).

	Subschalen			
	A	B	C	D
Inter-eenling (6 paren)	0,58 (47-70)	0,38 (19-66)	0,62 (50-83)	0,52 (35-68)
Inter-tweeling (3 paren)	0,74 (66-83)	0,57 (36-70)	0,78 (71-80)	0,68 (63-73)
Eenling-vierling (Winer-r)	0,83	0,65	0,90	0,78

De eerste regel van de tabel geeft per subschaal de gemiddelde coëfficiënt van de (tussen vier beoordelaars mogelijke) zes inter-eenling correlaties. Tussen haakjes zijn de uiterste waarden vermeld. De tweede regel geeft per subschaal de gemiddelde coëfficiënt van de (tussen vier beoordelaars mogelijke) drie inter-tweeling correlaties met tussen haakjes de uiterste waarden.

De cijfers in de derde rij zijn geen Pearson-coëfficiënten, maar coëfficiënten berekend volgens een formule van WINER (1962), die ook gebruikt is door BEYAERT (1966). Deze Winer-coëfficiënt wordt mede bepaald door de gemiddelde correlatie van een aantal eenlingen met de uit hen samengestelde meerling. In hoeverre deze coëfficiënt vergelijkbaar is met de Pearson-coëfficiënt, is mij overigens nog niet geheel duidelijk.

De cijfers voor de eindscores zijn als volgt:

Inter-eenling: 0,83 (76-87)

Inter-tweeling: 0,91 (87-96)

Eenling-vierling (Winer): 0,94.

Duidelijk blijkt uit deze cijfers, dat het gebruik van synthetische beoordelaars veel betere resultaten geeft dan het gebruik van in-

dividuele. Een tweeling is reeds aanmerkelijk trefzekerder dan een eenling. Een coëfficiënt van 0,94 is in de gedragswetenschappen bijzonder hoog.

Hoewel in mijn betoog het accent ligt op de interbeoordelaar-betrouwbaarheid, volgen thans nog enkele gegevens over de prestaties van synthetische beoordelaars op het gebied van het samenhangen-onderzoek. Wanneer synthetische beoordelaars in staat zijn, om beter dan individuele beoordelaars ondubbelzinnige verschijnselen te produceren, dan moeten zij ook samenhangen tussen verschijnselen duidelijker kunnen aangeven, hetgeen zich onder bepaalde omstandigheden moet uiten in hogere correlatiecoëfficiënten. Deze omstandigheden hebben zich blijkbaar voorgedaan bij de drie samenhangen, die wij hebben onderzocht, zoals blijkt uit tabel 2.

TABEL 2. De prestaties van individuele en synthetische beoordelaars uitgedrukt in validiteitscoëfficiënten verkregen bij een vergelijking van de WAV-eindscore met drie andere aanpassingsmaatstaven (N = 25).

	vier eenlingen	één vierling
PUK-schalen, totaalscore (tweeling)	0,86 (81-89)	0,92
Graffelschaal, eindscore (eenling)	0,72 (69-77)	0,77
Rangschikking naar aangepastheid (tweeling)	0,73 (68-76)	0,78

De hierin opgenomen coëfficiënten geven aan, wat de samenhang is tussen de WAV-eindscore enerzijds en de totaalscore van de PUK-schalen, de eindscore van de Graffelschaal en een rangschikking naar aangepastheid anderzijds.

Ter verkrijging van deze cijfers werd tijdens het reeds beschreven onderzoek met de WAV over dezelfde patiënten een beoordeling met de PUK-schalen en met de Graffelschaal uitgevoerd en werd een rangschikking naar aangepastheid gemaakt. De beoordeling met de PUK-schalen en het maken van de rangschikking geschiedde door twee parallel werkende verplegers, zodat met de verkregen uitkomsten tweelingen geconstrueerd konden worden. Met de Graffelschaal werd per patiënt slechts één beoordeling gemaakt door een groeps-leider in de arbeidstherapie.

In de eerste kolom van tabel 2 vindt men per maatstaf het gemiddelde van de vier correlatiecoëfficiënten, die worden verkregen,

wanneer de WAV door vier individuele beoordelaars wordt gehanteerd; tussen haakjes worden de uiterste waarden vermeld.

De tweede kolom geeft de correlatiecoëfficiënten, zoals die worden, wanneer men een vierling construeert.

Het is duidelijk, dat ook in dit samenhangen-onderzoek de synthetische beoordelaar meer heeft gepresteerd dan de individuele. De nuttigheid van deze superioriteit moge blijken uit b.v. het volgende: Wanneer de correlatie tussen de WAV-eindscore en de totaalscore van de PUK-schalen 0,86 bedraagt, dan kan men nog menen, dat ze niet helemaal hetzelfde verschijnsel aanduiden, maar bij een coëfficiënt van 0,92 begint het er erg op te lijken dat beide maatstaven praktisch hetzelfde meten.

CONCLUSIE

Op grond van onze gegevens lijkt mij het vermoeden gewettigd, dat de superioriteit van de synthetische beoordelaar meer is dan alleen maar een theoretisch interessante curiositeit. Naar mijn mening kan de synthetische beoordelaar in de dagelijkse inrichtingspraktijk een waardevolle bijdrage leveren en moet deze mogelijkheid meer dan thans het geval is worden benut. De praktische bezwaren behoeven niet groot te zijn; wanneer men eenmaal beschikt over schalen, is de constructie van synthetische beoordelaars verrassend eenvoudig.

SAMENVATTING

Door individuele beoordelingen te combineren worden „synthetische” beoordelaars gevormd, wier beoordelingen een hogere interbeoordelaar-betrouwbaarheid te zien geven en tot hogere validiteitscijfers kunnen leiden dan de beoordelingen van individuele beoordelaars. De mate, waarin de synthetische beoordelaar superieur is, werd gedemonstreerd aan de hand van gegevens, die werden verkregen met de experimentele Warnsveldse Aanpassingsschaal voor Verpleegafdelingen (WAV), de PUK-schalen en de Graffelschaal.

SUMMARY

By combining the behavior ratings from two or more evaluators it is possible to obtain "synthetic" ratings, which show higher interrater reliabilities and higher validities than individual ones. The extent of this superiority is demonstrated by reliability and validity coefficients obtained in an experiment in which four nurses rated 25 patients, using the Warnsveld Ward Adjustment Scale, a not yet fully developed scale of the Lucero and Meyer type. These ratings were compared with evaluations made by means of two existing scales.

LITERATUUR

- Beyaert, F. H. L. — 1966. *Gedragbeoordeling in het psychiatrisch ziekenhuis*, Leiden.
- Burger, A.W., Waard, J.A.A.de, Wilde, G.J.S. — 1966. *Voorlopige handleiding bij de Graffelschaal*, Amsterdam.
- Hofstätter, P. R. — 1957. *Gruppendynamik*, Hamburg.
- Jensen, M. B. and Morris, W. E. — 1960. Reliability-unreliability of ancillary psychiatric evaluations. *Journal of clinical psychology*, 16, 248-252.
- Winer, B. J. — 1962. *Statistical principles in experimental design*, London, pag. 128.

MEDEDELINGEN

Van 15 tot en met 26 augustus zal in het psychiatrisch centrum „St.Willibrord” te Heiloo een cursus worden georganiseerd die bedoeld is als een introductie in de theoretische en praktische aspecten van de *gedragstherapie*. Deze cursus wordt geleid door Prof. Dr. J. Wolpe (Temple University, Philadelphia).

In het kader van deze cursus zal Prof. Wolpe op de vrijdagen 19 en 26 augustus een reeks uitzendingen (zo mogelijk met patienten-demonstraties) geven die voor een groter gehoor toegankelijk zijn.

Een ieder die belangstelling heeft, kan zich *schriftelijk* opgeven bij ondergetekenden. Als onkostenvergoeding voor elk van deze vrijdagen is een bedrag vastgesteld van f 15,—. In dat bedrag zijn de kosten voor een eenvoudige lunch inbegrepen.

Aanmeldingen voor het volgen van de cursus in zijn geheel kunnen niet meer worden gehonoreerd. Het ligt echter in de bedoeling, dat de vereniging voor Gedragstherapie (secretariaat Mevr. Drs. B. J. Bakker-de Pree p.a. 2e Constantijn Huygensstraat 34, Amsterdam) het volgende jaar een soortgelijke cursus zal organiseren.

Dr. J. J. DIJKHUIS

Drs. F. A. PELGRIM

Psychiatrisch Centrum „St.Willibrord”, Heiloo

Van het Algemeen Nederlands Gezelschap voor de studie van zelfmoordproblemen (voorlopig secr.: Wilhelminapark 26, Utrecht. Telefoon 030 - 2 38 46) werd de volgende brief ontvangen ter publikatie.

Utrecht, 13 mei 1966

L.S.

Van 21 tot 25 september 1965 vond in Bazel het 3e Internationale Congres voor zelfmoordprophylaxe plaats, uitgaande van „The International Association for Suicide-Prevention”. Duidelijk bleek daar dat een bundeling van Nederlandse geïnteresseerden in het zelfmoordprobleem wenselijk was. Op 1 oktober 1965 werd daarom opgericht het Algemeen Nederlands Gezelschap voor de studie van zelfmoordproblemen. De bedoeling van dit gezelschap is:

1. te bundelen alle personen en instellingen die op wetenschappelijk of praktisch gebied zich geheel of gedeeltelijk bezighouden met - of geïnteresseerd zijn in - de verschillende facetten van het zelfmoordprobleem.