

De 'Nurses Observation Scale for Inpatient Evaluation' (NOSIE-30): interbeoordelaarsbetrouwbaarheid, stabieliteit en onderscheidend vermogen

door A. Hafkenscheid

Samenvatting

Van de 'Nurses Observation Scale for Inpatient Evaluation' (NOSIE-30) werden de interbeoordelaarsbetrouwbaarheid, stabieliteit en het onderscheidend vermogen onderzocht in een heterogene steekproef van opgenomen psychiatrische patiënten ($68 \leq N \leq 85$). Per afdeling beoordeelde een tweetal verpleegkundigen de daar opgenomen patiënten, onafhankelijk van elkaar. Totaal: zes beoordelaars. De interbeoordelaarsbetrouwbaarheden van de subschalen 'Sociale vaardigheid', 'Zelfverzorging' en 'Vertraagde motoriek', alsmede van de NOSIE-totaalscores, waren redelijk goed ($.71 \leq r \leq .84$). Binnen één van de drie paren beoordeelde de ene verpleegkundige patiënten echter consequent 'gunstiger' dan de andere. Voor de subschalen 'Sociale interesse', 'Prikkelbaarheid', 'Psychose' en 'Depressiviteit' waren de interbeoordelaarsbetrouwbaarheden matig tot laag voor tenminste één van de drie beoordelaarsparen.

Voor 5 van de 6 beoordelaars was de stabieliteit van de gedragsbeoordelingen – op subschaal en schaalniveau – in het algemeen hoog. De NOSIE bleek in staat patiënten met 'slechte' prognoses te onderscheiden van patiënten met 'gunstiger' prognoses.*

Inleiding

De 'Nurses Observation Scale for Inpatient Evaluation' (NOSIE-30) is een verpleegkundige observatieschaal voor gedragsbeoordeling van opgenomen psychiatrische patiënten. De NOSIE werd ruim 20 jaar geleden geïntroduceerd door Honigfeld, Gillis en Klett (1965, 1966). Inmiddels is een Nederlandse versie van de schaal beschikbaar. Deze versie is 'in public domain' (Dingemans 1985).

De NOSIE omvat 30 items, die verwijzen naar voor anderen waarneembare gedragingen ('Wordt gauw boos of geërgerd', 'Beweegt zich

* Karin Bekkers, Bart van Grinsven, Jeanien van Heuvel, Ton Kramer, Inge Snijders en Jan van Tamelen, psychiatrisch verpleegkundigen, komt dank toe voor hun gedragsbeoordelingen.

Automatische gegevensverwerking met behulp van de 'Superbrain'-computer (Intertec Data Systems), statistisch programma 'NWA STATPAK' (1982), Portland OR.

langzaam en traag'). Per item zijn er 5 antwoordmogelijkheden. Gedragingen worden beoordeeld naar de frequentie, waarmee ze binnen een afgebakende periode ('de afgelopen drie dagen') zijn vóórgelopen (met 'nooit' en 'altijd' als schaaluiteinden).

De NOSIE is multidimensioneel. De meetpretentie strekt zich uit tot de volgende gedragscategorieën (subschalen gelijknamig omschreven): 'sociale vaardigheid' ('social competence'), 'sociale interesse' ('social interest'), 'zelfverzorging' ('personal neatness'), 'prikkelbaarheid' ('irritability'), 'manifeste psychose' ('manifest psychosis'), 'motorische vertraging' ('retardation') en 'depressiviteit' ('depression').

De NOSIE vindt onder andere toepassing als effectmaat in farmacotherapeutisch onderzoek (zie bij voorbeeld Eccleston, Fairbairn, Hasanyeh, McClelland en Stephens 1985; De Cuyper, Bollen, Van Praag en Verstraeten 1986). Sinds de introductie van de NOSIE is er – in het anglo-Amerikaanse taalgebied – psychometrisch ruimschoots ervaring mee opgedaan. Dingemans, Bleeker, Bakker-Winnubst en Frohnde Winter (1983) vermelden 7 publikaties, waarin betrouwbaarheids- en validiteitsgegevens worden gepresenteerd. Van de Nederlandse versie onderzochten deze auteurs de dimensionele structuur en interbeoordelaarsbetrouwbaarheid in een steekproef van kortdurend opgenomen psychiatrische patiënten (Dingemans e.a. 1984); alsook de vergelijkbaarheid met een andere verpleegkundige observatieschaal – de 'Beoordelingsschaal voor Oudere Patiënten' (BOP; Van der Kam, Mol en Wimmers 1971) in een psychogeriatrische steekproef (Dingemans e.a. 1983). Onder andere bleek de dimensionele structuur (behoudens een transcultureel verschil: items uit de 'sociale vaardigheids'- en 'zelfverzorgings'-subschaal bleken dimensioneel moeilijk onderscheidbaar) in belangrijke mate overeen te komen met de Amerikaanse. De interbeoordelaarsbetrouwbaarheid op subschaalniveau bleek in beide onderzoeken in het algemeen bevredigend.

Buiten deze twee onderzoeken is er – te oordelen naar gepubliceerd instrumenteel onderzoek – nog betrekkelijk weinig bekend over de psychometrische eigenschappen van de Nederlandstalige NOSIE. In het overzichtsboek van Visser, Van Vliet-Mulder, Evers en Ter Laak (1982) komt de NOSIE niet voor; evenmin in de latere aanvullingen op deze 'Documentatie van tests en testresearch in Nederland'. In het Sinai-Centrum, psychiatrisch ziekenhuis van de 'Joodse Geestelijke Gezondheidszorg', zijn enkele psychometrische kenmerken van de NOSIE onderzocht in een groep van 85 opgenomen psychiatrische patiënten. Met deze publikatie van de bevindingen wordt beoogd een bijdrage te leveren aan de accumulatie van psychometrische kennis omtrent de 'Nederlandse NOSIE'.

Methodes

Het Sinai-Centrum telt 3 psychiatrische afdelingen: een voor manne-

lijke en twee voor vrouwelijke patiënten. De NOSIE werd voor alle patiënten, die op de data van de gegevensverzameling aanwezig waren, ingevuld. Elke patiënt werd beoordeeld door twee gediplomeerde psychiatrische verpleegkundigen, die tenminste één jaar aan een afdeling verbonden waren. De beoordelaars vulden de schaal onafhankelijk van elkaar in. In totaal waren 6 verpleegkundigen – 2 per afdeling – bij de beoordeling betrokken. De beoordelaars waren van de bedoeling van het onderzoek – evaluatie van het meetinstrument en niet van hun persoonlijke observatievaardigheid of iets dergelijks – op de hoogte. De beoordeling vond herhaald plaats, met een tussenperiode van een week.

De eerste meting ($N = 85$) diende ter bepaling van de *interbeoordelaarsbetrouwbaarheid*. De patiënten die een week na de eerste beoordeling nog waren opgenomen ($N = 68$) werden door dezelfde twee verpleegkundigen beoordeeld. Met behulp van deze tweede meting werd een indruk verkregen van de *stabiliteit* van de gedragsbeoordelingen. Om een indicatie te krijgen van het *onderscheidend vermogen* van de NOSIE werd onder andere gebruik gemaakt van prognostische indelingen, die beide aan de drie afdelingen verbonden psychiaters en hoofdverpleegkundigen voor een ander doel – een intern beleidsonderzoek naar chroniciteit (Hafkenscheid 1986) – aanbrachten ($N = 73$).

Besloten werd de dimensionele indeling van Dingemans e.a. (1984) aan te houden. De 'sociale vaardigheids'- en 'zelfverzorgings'-subschalen werden echter niet samengevoegd. Alhoewel deze subschalen in het onderzoek van juistgenoemde auteurs empirisch niet te onderscheiden waren, zou 'samenvoeging vooraf' de eventuele onderscheidbaarheid van beide subschalen in deze groep reeds bij voorbaat – kunstmatig – verhullen.

Proefpersonen

Geslacht: 54 vrouwelijke, 31 mannelijke patiënten, leeftijd: $m = 53.8$ jaar; $sd = 17.7$; leeftijdsverdeling: 10 patiënten in de leeftijdsgroep '25-34 jaar' (11.7%); 21 (24.7%): 35-44 jaar; 16 (18.8%): 45-54 jaar; 11 (12.9%): 55-64 jaar; 20 (23.5%): 65-74 jaar; 6 (7%): 75-84 jaar; 1 (1.1%): 85-94 jaar.

Opnameduur: 42 patiënten (49.4%) zijn minder dan een jaar opgenomen, van wie: 12 minder dan een maand; 7: 1-3 maanden; 16: 3-6 maanden; 4: 6-9 maanden; 3: 9-12 maanden. Van de 43 (50.6%) patiënten die langer dan 1 jaar zijn opgenomen is de opnameduur voor 12 van hen '1-2 jaar'; voor eveneens 12 patiënten: 2-5 jaar; voor 19 patiënten: 5 jaar of langer. Opnamediagnoses volgens ICD-9-CM-classificatie: organische psychotische aandoeningen (290-294): 3 (3.5%), overige psychosen (295-299): 54 (63.5%), neurotische stoornissen, persoonlijkheidsstoornissen en overige niet-psychische stoornissen (300-316): 28 (32.9%).

Resultaten

Interbeoordelaarsbetrouwbaarheid – De interbeoordelaarsbetrouwbaarheid werd berekend per beoordelingspaar, voor de afdelingen afzonderlijk. De drie paren beoordeelden een bij benadering gelijk aantal patiënten: de beoordelaars van afdeling A 29 patiënten, van afdeling B 30 patiënten, van afdeling C 26 patiënten. Op *item-niveau* werd de interbeoordelaarsbetrouwbaarheid berekend met behulp van Cohens 'kappa': k (zie bij voorbeeld Landis en Koch 1975). De interbeoordelaarsbetrouwbaarheid blijkt voor alle 30 items, zij het in verschillende mate, laag (zie appendix 1). Dat geldt voor alle drie beoordelaarsparen.

Alle beoordelingen per paar tegelijkertijd in ogenschouw genomen (dus 'kappa' berekend over alle 30 items): afdeling A: $k = .33$; afdeling B: $k = .26$; afdeling C: $k = .27$. Met 'kappa' wordt strikte overeenstemming tussen beoordelaars berekend. Niet elk beoordelingsverschil is uiteraard even ernstig (bij voorbeeld: antwoordmogelijkheden als 'gewoonlijk' en 'altijd' verschillen slechts in nuance). Daarom werden – over alle beoordelingen per paar – eveneens 'gewogen kappa's' berekend. Gewichten werden toegekend volgens de formule van Cicchetti (Landis en Koch 1975). Voor afdeling A: $k_w = .51$, voor afdeling B: $k_w = .43$, voor afdeling C: $k_w = .44$, waarbij $k_w =$ 'gewogen kappa'. Ook de 'gewogen kappa's' zijn laag.

De interbeoordelaarsbetrouwbaarheden op *subschaal-* en *schaalniveau*, berekend met behulp van PM-correlaties, staan weergegeven in tabel 1.

Tabel 1: Interbeoordelaarsbetrouwbaarheid NOSIE-schalen

	Afdeling			Totaal
	A	B	C	
Sociale vaardigheid	.77	.82	.75	.68
Sociale interesse	.76	.66	.78	.64
Zelfverzorging	.83	.71	.80	.79
Prikkelbaarheid	.60	.64	.70	.54
Psychose	.88	.57	.68	.75
Vertraagde motoriek	.84	.73	.82	.65
Depressiviteit	.71	.46	.64	.38
NOSIE-30: totaalscores	.83	.76	.79	.70

Voor alle drie afdelingen is de overeenstemming tussen beoordelaars op de subschalen 'Sociale vaardigheid', 'Zelfverzorging' en 'Vertraagde motoriek' als 'redelijk goed' (in navolging van de kwalificaties van Van der Kam e.a. (1971)) te kenmerken. Dat geldt ook voor de NOSIE-totaalscores. De interbeoordelaarsbetrouwbaarheid voor de subschaal 'Sociale interesse' is bij één beoordelaarspaar (afdeling B) wat lager. Voor tenminste twee beoordelaarsparen is de overeenstemming op de

subschalen 'Prikkelbaarheid', 'Psychose' en 'Depressiviteit' matig tot laag. Rechts in tabel 1 staan de interbeoordelaarsbetrouwbaarheden per (sub)schaal voor alle 6 beoordelaars (anders gesteld: voor alle 85 beoordeelde patiënten) te zamen weergegeven. De lage overeenstemming tussen de beoordelaars van afdeling B op de subschaal 'Depressiviteit' werkt door in de lage overeenstemming op deze subschaal, voor alle 6 beoordelaars te zamen.

Hoge correlatiecoëfficiënten impliceren slechts in een bepaalde zin overeenstemming tussen beoordelaars. Als beoordelaars *systematisch* verschillen in hun beoordelingen, kan een hoge *samenhang* tussen (sub)schaalscores bereikt worden, tegelijkertijd met (eventueel aanzienlijke) verschillen in de *hoogte* van die scores. Daarom werden – per beoordelaarspaar – voor elke subschaal, alsmede voor de NOSIE-totaalscores, beoordelaarsverschillen in die zin getoetst met de t-toets voor gepaarde waarnemingen (zie appendix 2).

De beoordelaars van afdeling A verschillen op twee subschalen statistisch significant; de beoordelaars van afdeling B op drie subschalen; de beoordelaars van afdeling C op vijf subschalen, alsmede op de NOSIE-totaalscores. De beoordelaars van afdeling C verschillen op alle subschalen consequent in hun gedragsbeoordelingen: de ene beoordelaar beoordeelt patiënten systematisch 'gunstiger' dan de andere. Met de maat voor 'Effect Grootte' ('Effect Size', gesymboliseerd met 'd') van Cohen (1977) kan de betekenis van – statistisch significante – verschillen worden weergegeven. In termen van 'Effect Grootte' (voor gepaarde waarnemingen): bij de beoordelaars van afdeling A is één verschil (subschaal 'Prikkelbaarheid') 'middelgroot' ($.50 \leq d < .80$), bij de beoordelaars van afdeling B is één verschil 'middelgroot' (subschaal 'Sociale interesse') en één 'groot' ($d \geq .80$; subschaal 'Depressiviteit'), bij de beoordelaars van afdeling C zijn vier verschillen 'middelgroot' (subschalen 'Sociale interesse', 'Prikkelbaarheid', 'Psychose' en 'Depressiviteit') en drie 'groot' ('Sociale vaardigheid', 'Vertraagde motoriek' en de NOSIE-totaalscores).

De overige getoetste verschillen hebben – in termen van 'Effect Grootte' – weinig te betekenen.

Stabiliteit – Voor alle 6 beoordelaars afzonderlijk werd de stabiliteit – de *temporele* betrouwbaarheid – van de gedragsbeoordelingen berekend met PM-correlaties. De stabiliteitscoëfficiënten staan in tabel 2 weergegeven (tussenperiode: 1 week).

Met name voor de beoordelaars van de afdelingen A en C zijn de stabiliteitscoëfficiënten vrijwel zonder uitzondering hoog. Voor de beoordelaars van afdeling B is dat – zeker voor beoordelaar 3 – anders. Dat betekent overigens niet automatisch, dat de beoordelaars van afdeling B 'slechter' (want: minder stabiel) observeren. Afdeling B is een opnameafdeling, waar verhoudingsgewijs veruit de meeste kortdurend opgenomen patiënten verblijven. Bovendien is het verloop van patiënten

op deze afdeling het grootst. Patiënten die relatief kort zijn opgenomen zijn bij beoordelaars doorgaans het minst bekend. Daarnaast doen zich in de beginfase van een psychiatrische behandeling in korte tijd soms tamelijk ingrijpende gedragsveranderingen voor, die in de beoordelingen adequaat tot uiting kunnen komen.

Tabel 2: Stabiliteit NOSIE-schalen

Beoordelaar	Afdeling					
	A		B		C	
	1	2	3	4	5	6
Sociale vaardigheid	.94	.87	.71	.84	.81	.86
Sociale interesse	.85	.76	.50	.72	.75	.82
Zelfverzorging	.94	.94	.82	.80	.90	.94
Prikkelbaarheid	.83	.76	.60	.87	.93	.83
Psychose	.93	.96	.60	.71	.85	.72
Vertraagde motoriek	.84	.84	.38	.51	.79	.84
Depressiviteit	.85	.80	.80	.57	.59	.90
NOSIE-30: totaalscores	.95	.93	.43	.70	.82	.88
	n = 29		n = 15		n = 24	

Samenhang tussen subschalen – Voor de – per beoordelaarspaar – gesommeerde subschaalscores werden de intercorrelaties tussen de subschalen berekend. Deze staan weergegeven in tabel 3.

Tabel 3: Intercorrelaties (per beoordelaarspaar) gesommeerde schaalscores

		S.v.	S.i.	Ze	Pr	Ps	V.m.	De
Sociale vaardigheid	(S.v.)	1.00	.69	.81	.68	.56	.77	.34
Sociale interesse	(S.i.)		1.00	.50	.38	.30	.74	.15
Zelfverzorging	(Ze)			1.00	.48	.54	.60	.17
Prikkelbaarheid	(Pr)				1.00	.50	.29	.34
Psychose	(Ps)					1.00	.26	-.04
Vertraagde motoriek	(V.m.)						1.00	.27
Depressiviteit	(De)							1.00

De samenhang tussen subschalen is soms aanzienlijk. Die samenhang is gedeeltelijk artificieel: sommige items (met name die uit de 'Sociale vaardigheids'-subschaal) keren in meerdere subschalen terug. De items uit de 'Sociale vaardigheids'- en 'Zelfverzorgings'-subschalen vertonen geen 'overlap'. Niettemin correleren deze subschalen onderling hoog ($r = .81$; tevens de hoogste intercorrelatie). Dit bevestigt de bevinding van Dingemans e.a. (1983, 1984), dat Nederlandse verpleegkundigen beide gedragsdimensies *in hun beoordelingen* niet duidelijk onderscheiden (wat niet wil zeggen, dat zij beide dimensies *conceptueel* niet zouden onderscheiden).

Onderscheidend vermogen – De validiteit van de NOSIE wordt mede bepaald door de mogelijkheid ‘ernstig gestoorde’ (‘ernstig geïnvaleerde’) van ‘minder ernstig gestoorde’ (‘tijdelijk of minder ernstig geïnvaleerde’) patiënten te onderscheiden. Vier prognostische tweedelingen van beide afdelingspsychiaters en hoofdverpleegkundigen dienden als criteriummaten. Zij hadden geen toegang tot de NOSIE-scores. Het onderscheidend vermogen werd berekend over de – per beoordelaarspaar – gesommeerde NOSIE-totaalscores. Vooraf werd getoetst in hoeverre beide paren afdelingspsychiaters/hoofdverpleegkundigen in hun toewijzing verschillen. Dat geschiedde voor elk van de vier tweedelingen (zie appendix 3). Op geen van de vier criteriummaten zijn er statistisch significante toewijzingsverschillen. In tabel 4 staan de gemiddelden en spreidingen voor de vier tweedelingen weergegeven. Hogere gemiddelden duiden op ernstiger gestoordheid.

Tabel 4: NOSIE-totaalscores bij prognostische tweedelingen

	Ja	Nee
Verblijfspatiënt	m = 91.7 sd = 26.1 n = 35	m = 60.4 sd = 29.2 n = 38
Opname t/m 31/12 geïndiceerd	m = 81.98 sd = 30.3 n = 50	m = 61.2 sd = 30.4 n = 23
Beschermd wonen mogelijk	m = 54.04 sd = 19.8 n = 23	m = 85.3 sd = 31.5 n = 50
Zelfstandigheid haalbaar	m = 49 sd = 29.3 n = 16	m = 82.8 sd = 28.4 n = 57

Voor elk van de vier tweedelingen is het verschil tussen de prognostisch gunstiger en de prognostisch minder gunstige subgroep statistisch significant in de verwachte richting (zie appendix 4). In termen van ‘Effect Grootte’ zijn de verschillen voor de criteria ‘Verblijfspatiënt’, ‘Beschermd wonen momenteel mogelijk’ en ‘Volledige zelfstandigheid na ontslag haalbaar’ als ‘groot’ ($d \geq .80$) aan te merken; voor ‘Opname t/m 31/12 geïndiceerd’ als ‘middelgroot’ ($.50 \leq d < .80$). Bij het criterium ‘Beschermd wonen momenteel mogelijk’ dient de kwalificatie in termen van ‘Effect Grootte’ met voorzichtigheid te worden gehanteerd: de subgroepen verschillen daar niet alleen aanzienlijk in steekproefomvang, maar eveneens in steekproefspreiding (waarmee de *populatiespreiding* wordt geschat). In die omstandigheid kan de d-maat aan forse vertekening onderhevig zijn (Cohen 1977).

Conclusies

De NOSIE is – met zekere restricties – een bruikbaar instrument ten behoeve van de gedragsbeoordeling van opgenomen psychiatrische patiënten gebleken.

Op *item*-niveau bleek de interbeoordelaarsbetrouwbaarheid van de NOSIE sterk onvoldoende. Hoewel dit aan de bruikbaarheid van de schaal geen wezenlijke afbreuk doet, pleit dit er in elk geval tegen gedragsbeoordelingen op het niveau van afzonderlijke items te interpreteren.

Op *subschaal*-niveau bleek de interbeoordelaarsbetrouwbaarheid voor drie van de zeven subschalen adequaat voor alle drie beoordelaarsparen. Dat geldt eveneens voor de NOSIE-totaalscores.

De interbeoordelaarsbetrouwbaarheden van de subschalen zijn overigens vrijwel zonder uitzondering lager dan de zeer hoge waarden die door Dingemans e.a. (1983, 1984) zijn gerapporteerd. Of dit een gevolg is van het feit dat onze steekproef aanzienlijk kleiner is dan de steekproeven van deze onderzoekers (en derhalve gevoeliger voor toevalsfluctuaties) is moeilijk na te gaan. Want ook de *berekeningswijze* verschilt. Dingemans e.a. berekenden interbeoordelaarsbetrouwbaarheden op grond van 'factor matching', een methode die per definitie aan kanskapitalisatie onderhevig is (Ten Berge 1980). In onderhavig onderzoek zijn de interbeoordelaarsbetrouwbaarheden simpelweg gebaseerd op correlaties tussen *ongewogen* subschaalscores, en zijn daarvoor wellicht minder geflatteerd.

De *stabiliteit* van de gedragsbeoordelingen op (sub)schaalniveau bleek in het algemeen – voor vijf van de zes beoordelaars – hoog. Het onderscheidend vermogen van de NOSIE kon in dit onderzoek worden bevestigd.

Wat betreft het gebruik van de NOSIE als aanvulling op verpleegkundige rapportage of als effectmaat in 'outcome'-onderzoek geven onze bevindingen aanleiding tot het plaatsen van enkele kanttekeningen. In de eerste plaats zijn de restricties, die voor de NOSIE op *item*-niveau gelden, eveneens van toepassing op een aantal subschalen: lage ('Depressiviteit') of matige interbeoordelaarsbetrouwbaarheden, die de interpretatie van scores of scoreveranderingen bemoeilijken. In de tweede plaats kunnen – ook bij redelijk goede interbeoordelaarsbetrouwbaarheden – aanzienlijke verschillen tussen beoordelaars optreden: verschillen in de neiging, gedrag als gestoord te beoordelen, los van specifieke patiënten. De interpreteerbaarheid van gedragsbeoordelingen is daarmee afhankelijk van kennis omtrent specifiek beoordelaarsgedrag. Om eventuele beoordelaarsverschillen althans zichtbaar te maken, is het raadzaam – ook bij gebruik binnen de kliniek – patiënten door (ten minste) twee verpleegkundigen 'gelijktijdig' te laten beoordelen. In navolging van Honigfeld e.a. (1965, 1966) zouden de scores van de afzonderlijke beoordelaars vervolgens kunnen worden gesommeerd.

Wanneer de NOSIE als maat voor gedragsverandering wordt gehanteerd, is het – vanwege eventuele beoordelaarsverschillen als ‘foutenbron’ – in elk geval aan te bevelen op de verschillende meetmomenten van dezelfde beoordelaar(s) gebruik te maken. Dan kan er vanzelfsprekend niet mee worden volstaan begin- en vervolgcores simpelweg van elkaar af te trekken. Immers: scoreverandering impliceert niet noodzakelijk ‘werkelijke’ gedragsverandering. Scores fluctueren ook in zekere mate vanwege onbetrouwbaarheid in de beoordelingen zelf.

Met hun index voor ‘significant change’ stellen Christensen en Mendoza (1986) voor ‘ruwe’ scoreverschillen te relateren aan een schatting van de standaardmeetfout voor verschillcores (s_{diff}). Zodoende kan bepaald worden of scoreveranderingen onbetrouwbaarheidsmarges overstijgen. Over de NOSIE-totaalscores is deze index berekend voor elk van de zes beoordelaars afzonderlijk; daarnaast voor twee beoordelaars (uit hetzelfde beoordelaarspaar), die op verschillend tijdstip (tussenliggende periode: 1 week) dezelfde patiënten beoordelen (zie appendix 5). Zijn deze schattingen enigszins representatief, dan betekent dit dat (bij voorbeeld voor een – symmetrisch – 95%-betrouwbaarheidsinterval: $|x_2 - x_1| > 2 * s_{diff}$) pas bij vrij forse scoreveranderingen mag worden aangenomen, dat ook ‘werkelijke’ gedragsverandering heeft plaatsgevonden.

Literatuur

- Berge, J.M.F. ten (1980), *Faktoranalyse*. Heymans Bulletins Psychologische Instituut Rijksuniversiteit Groningen (HB-77-338-EX), Groningen.
- Christensen, L., en J.L. Mendoza (1986), A method of assessing change in a single subject: an alteration of the RC-Index. *Behavior Therapy*, 1, 305-308.
- Cohen, J. (1977), *Statistical power analysis for the behavioral sciences*. Academic Press, New York.
- Cuyper, H. de, J. Bollen, H.M. van Praag en D. Verstraeten (1986), Pharmacokinetics and therapeutic efficacy of haloperidol decanoate after loading dose administration. *British Journal of Psychiatry*, 148, 560-566.
- Dingemans, P.M.A.J., J.A.C. Bleeker, M. Bakker-Winnubst en M.L. Frohn-de Winter (1983), De NOSIE en de BOP nader onderzocht: een vergelijking tussen twee gedragsobservatieschalen in de psychogeriatric. *Tijdschrift Gerontologie en Geriatric*, 14, 223-230.
- Dingemans, P.M., J.A.C. Bleeker en M.L. Frohn-de Winter (1984), A cross-cultural study of the reliability and factorial dimensions of the Nurses' Observation Scale for Inpatient Evaluation (NOSIE). *Journal of Clinical Psychology*, 40, 169-172.
- Dingemans, P., brief d.d. 27/3/1985.
- Eccleston, D., A.F. Fairbairn, F. Hassanyeh, H.A. McClelland en D.A. Stephens (1985), The effect of propranolol and thioridazine on positive and negative symptoms of schizophrenia. *British Journal of Psychiatry*, 147, 623-630.
- Hafkenscheid, A. (1986), *Chroniciteit op de psychiatrische afdelingen van het Sinai-centrum: enkele getalsmatige indicaties*. Intern rapport, Amersfoort.

- Honigfeld, G., en C.J. Klett (1965), The Nurses' Observation Scale for Inpatient Evaluation. *Journal of Clinical Psychology*, 21, 65-71.
- Honigfeld, G., R.C. Gillis en C.J. Klett (1966), NOSIE-30: a treatment sensitive ward behavior scale. *Psychological Reports*, 19, 180-192.
- Kam, P. van der, F. Mol en M.F.H.G. Wimmers (1971), *Beoordelingsschaal voor Oudere Patiënten*. Van Loghum Slaterus, Deventer.
- Landis, J.R., en G.G. Koch (1975), A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). *Statistica Neerlandica*, 29, 151-161.
- Visser, R.S.H., J.C. van Vliet-Mulder, A. Evers en J. ter Laak (1982), *Documentatie van tests en testresearch in Nederland - 1982*. Nederlands Instituut van Psychologen, Amsterdam.

Appendix

1. Interbeoordelaarsbetrouwbaarheid op item-niveau

Van de 90 'kappa'-waarden ($3 \times 30 = 90$; namelijk: 3 beoordelaarsparen, 30 items): $.00 \leq k \leq .10$: 21; $.11 \leq k \leq .20$: 25; $.21 \leq k \leq .30$: 16; $.31 \leq k \leq .40$: 14; $.41 \leq k \leq .50$: 12; $.51 \leq k \leq .55$: 2.

Onder de 30 items zijn er geen die systematisch - door alle drie paren - betrouwbaarder beoordeeld worden dan andere. Lage en iets hogere 'kappa's' zijn ongeveer gelijkmatig over de beoordelaarsparen verdeeld.

2. Beoordelaarsverschillen in de hoogte van toegekende (sub)schaalscores

Met $\alpha = .02$ (tweezijdig getoetst) verschillen de (sub)schaalscores voor de beoordelaars van afdeling A op de subschalen 'Prikkelbaarheid' ($t = 3.2$; d.f. = 28) en 'Psychose' ($t = 2.4$; d.f. = 28) statistisch significant; de beoordelaars van afdeling B op de subschalen 'Sociale interesse' ($t = 3.8$; d.f. = 29), 'Vertraagde motoriek' ($t = 1.96$; d.f. = 29) en 'Depressiviteit' ($t = 5.99$; d.f. = 29); de beoordelaars van afdeling C op de subschalen 'Sociale vaardigheid' ($t = 4.8$; d.f. = 25), 'Sociale interesse' ($t = 3.6$; d.f. = 25), 'Prikkelbaarheid' ($t = 3.8$; d.f. = 25), 'Vertraagde motoriek' ($t = 6.3$; d.f. = 25), 'Depressiviteit' ($t = 3.1$; d.f. = 25), alsmede op de NOSIE-totaalscores ($t = 5.7$; d.f. = 25).

3. Toewijzingsverschillen afdelingspsychiaters/hoofdverpleegkundigen criteriummaten

Getoetst met Chi-kwadraat (herberekend met 'Yates correction') en $\alpha < .01$: $\chi_c^2 = 2.98$ (verblijfspatiënt), $\chi_c^2 = .28$ (opname t/m 31/12 geïndiceerd), $\chi_c^2 = 1.08$ (beschermde wonen mogelijk), $\chi_c^2 = .12$ (volledige zelfstandigheid haalbaar), d.f. = 1: geen statistisch significante toewijzingsverschillen.

4. Verschillen NOSIE-totaalscores bij vier prognostische tweedelingen

Met de t-toets voor onafhankelijke steekproeven (eenzijdig getoetst, $\alpha < .005$; d.f. = 71): 'Verblijfspatiënt': $t = 4.77$; 'Opname t/m 31/12 geïndiceerd': $t = 2.67$; 'Beschermd wonen mogelijk': $t = 5.06$; 'Volledige zelfstandigheid haalbaar': $t = 3.996$.

5. Waarden 'index voor significant change'

Berekend over de NOSIE-totaalscores van elk van de zes beoordelaars afzonderlijk: $7.3 \leq s_{\text{diff}} \leq 8.4$ voor 4 beoordelaars; $s_{\text{diff}} \approx 11$ (afgerond) voor 2 beoordelaars. Wanneer de NOSIE-totaalscores op twee meetmomenten van twee *verschillende* beoordelaars afkomstig zijn: $10.7 \leq s_{\text{diff}} \leq 14.3$, geschat voor elk van de drie beoordelaarsparen.

Schrijver is als psycholoog verbonden aan het Sinai-centrum van de 'Joodse Geestelijke Gezondheidszorg', Amersfoort.