

Betrouwbaarheidsaspecten van de Present State Examination in vier verschillende populaties

door T.J. Duine, P.P.G. Hodiamont, P.G.M. Peer
en L.H. Boerma

Inleiding

Tot het scala van problemen waarvoor men zich gesteld ziet als men binnen de psychiatrie diagnostiek wil bedrijven (Hodiamont, 1983), behoort ongetwijfeld het vinden van consensus tussen verschillende deskundigen met betrekking tot het constateren van symptomen bij een te diagnosticeren individu. Het onderhavige artikel houdt zich bezig met de vraag naar overeenstemming tussen beoordelaars bij het toepassen van een specifiek diagnostisch instrument, te weten de Present State Examination (PSE).

Te oordelen naar het aantal publikaties neemt de PSE binnen de psychiatrische diagnostiek een steeds belangrijker plaats in. Nadat het instrument in de Medical Research Council van het Institute of Psychiatry te Londen was ontwikkeld (Winge.a., 1967), is het inmiddels in vele landen vertaald en toegepast, in met name klinische onderzoeks-situaties. De PSE heeft deze populariteit vooral te danken aan zijn opzet als semigestructureerd interview, waardoor interpretatieverschillen tussen beoordelaars geminimaliseerd kunnen worden.

Het interview is gebaseerd op de traditioneel medische benadering, waarin de onderzoeker met het oog op een bepaald symptoom aan de patiënt een serie vragen stelt, die elk afhangen van het antwoord op een vorige vraag en pas worden beëindigd wanneer zekerheid bestaat over het symptoom. De onderzoeker heeft de mogelijkheid zich aan te passen aan de interviewsituatie of aan de reacties van de ondervraagde, zonder dat dit leidt tot een verlies van standaardisatie. Wel dient de onderzoeker een grondige training achter de rug te hebben, niet alleen

De schrijvers zijn verbonden aan het Regioproject Nijmegen, Verlengde Groenestraat 75, 6525 EJ Nijmegen. T.J. Duine (pedagoge), P.G.M. Peer (statistica) en L.H. Boerma (sociologe) zijn werkzaam op het Instituut Sociale Geneeskunde, P.P.G. Hodiamont (zenuwarts) is stafmedewerker Sociale Psychiatrie, allen Katholieke Universiteit Nijmegen.

in interviewtechniek, maar ook in de opvatting en het hanteren van symptoomdefinities. Voor een uitvoerige beschrijving van de PSE en de verwerking ervan zij verwezen naar Verhey (1981).

Het gestandaardiseerde karakter van het interview maakt het mogelijk de betrouwbaarheid vast te stellen. Dit is ook diverse malen gedaan (o.a. Kendell e.a., 1968; Pen e.a., 1977 en Wing e.a., 1967). In de meeste van deze studies gaat het om klinische populaties. Er zijn echter ook toepassingen op niet-klinische psychiatrische patiënten (Cooper, 1977) en zelfs op een 'normale' populatie (Wing, 1977).

Het onderhavige artikel beschrijft een onderzoek, dat voorafging aan een groot regionaal onderzoek naar gezondheid en gezondheidsgedrag in de regio Nijmegen. Het hoofddoel van dit project is inzicht te krijgen in het ziekte- en hulpzoekgedrag van de bevolking in de gezondheidsregio Nijmegen als gevolg van somatische klachten, psychiatrische symptomen en psychosociale problemen en waar mogelijk in de adequaatheid ervan.

Voor de vaststelling van psychiatrische symptomen is gebruik gemaakt van de PSE. Tien interviewers zijn geselecteerd om na een grondige training bij naar schatting 500 personen de PSE af te nemen. Doel van het vooronderzoek was het meten van de betrouwbaarheid tussen de interviewers onderling (inter-rater reliability). Bovendien bood een dergelijk vooronderzoek de mogelijkheid de betrouwbaarheid van het instrument te testen voor verschillende populaties. We kunnen de probleemstelling van het artikel als volgt formuleren:

1. Hoe groot is de overeenkomst tussen onafhankelijke beoordelaars van de PSE-symptomen wanneer beiden na elkaar een live-interview afnemen bij dezelfde patiënt (live-live opzet) en wanneer een live-interview van een beoordelaar wordt nagescoord door een andere beoordelaar (live-tape opzet).

2. In hoeverre levert afname van de PSE bij verschillende populaties verschillen op in de betrouwbaarheid van het instrument. Het gaat hierbij om 4 groepen die vermoedelijk duidelijk van elkaar te onderscheiden zijn op basis van psychiatrische problematiek.

De onderzoeksopzet

De vier categorieën zijn als volgt verdeeld over het design:

categorie	aantal	aantal vergelijkingen	
	personen	live-live	live-tape
niet-patiënten	20	20	20x2
huisartspatiënten	10		10
APZ-patiënten	20	20	20x2
patiënten van een neurosekliek	10		10

De verdeling van de geïnterviewde personen over de 4 categorieën is voor alle interviewers hetzelfde.

De categorieën-

1. De niet-patiëntengroep bestaat uit 20 personen, die op het moment van het interview en de 3 maanden daarvoor niet in behandeling waren bij huisarts, specialist, kliniek of andere hulpverlenende instantie.

We hebben deze groep gekozen om een zo groot mogelijk contrast te kunnen krijgen met de andere categorieën.

2. De huisartsencategorie bestaat uit 10 patiënten van een gezondheidscentrum, die in de week van het interview een afspraak hadden met hun huisarts. Na dit consult werd hen gevraagd of zij wilden deelnemen aan het vooronderzoek. Zij werden dus niet geselecteerd op bepaalde klachten, het ging alleen om het feit, dat zij die week hun huisarts raadpleegden.

Hoewel het ook hier een niet-klinische patiëntencategorie betreft, valt te verwachten dat in deze groep meer symptomatologie aanwezig zal zijn dan in de eerste groep.

3. De psychiatrische categorie wordt gevormd door in totaal 20 patiënten uit twee Algemeen Psychiatrische Ziekenhuizen (APZ). Het betreft veelal mensen die al langere tijd in de inrichting verblijven. Omtrent de aard van hun stoornis valt geen eenduidig antwoord te geven.

4. De laatste groep bestaat uit 10 patiënten van een neurosekliek. Mensen met psychotische symptomen worden in deze kliniek niet opgenomen of doorverwezen naar andere instellingen. We mogen dus aannemen dat we hier overwegend met neurotische symptomen te maken krijgen.

Design – In ons onderzoek is zowel het live-live als het live-tape design toegepast. Bij het live-tape design is een geluidsband van het live-interview afgeluisterd en gescoord door een tweede interviewer. Bij het live-live design zijn twee live-interviews afgenomen bij dezelfde respondent. De gemiddelde tijdsduur tussen twee live-interviews was 5 dagen. Deze combinatie van designs is zeer gebruikelijk bij betrouwbaarheidsonderzoek van de PSE (Kendell, 1968; Pen, 1977; Slooff, 1983). Toepassing van dit design leidt ertoe dat de resultaten kunnen worden vergeleken met die van andere studies. De opzet heeft echter nog andere voordelen.

Het live-tape design biedt de mogelijkheid om twee personen hetzelfde interview te laten beoordelen, zodat een belangrijke foutenbron wordt uitgesloten. Beide beoordelaars gaan immers uit van dezelfde steekproef van gedrag. Als nadeel geldt dat hiermee ook een bias kan optreden van de observator in de richting van de interviewer. De interviewer bepaalt namelijk het verloop van het interview, beslist

wanneer afsluitpunten gebruikt worden en heeft de vrijheid om alternatieve vragen te stellen. Dit probleem speelt niet bij de live-live-situatie. Beide interviewers nemen zelfstandig het interview af en kunnen elkaar dus niet beïnvloeden. Hier doemt echter een andere foutenbron op, namelijk het toestandbeeld van de ondervraagde. Dit kan veranderd zijn tussen het eerste en tweede interview, of de proefpersoon kan van mening zijn dat hij bepaalde zaken in het eerste interview al zo uitvoerig heeft verteld, dat een tweede keer kan worden volstaan met een beknopter antwoord.

Bij een combinatie van beide designs kunnen de nadelen die aan elk afzonderlijk kleven, elkaar enigszins relativeren. Helaas was deze optimale situatie niet bij alle groepen te bereiken. Om financiële en technische redenen moest de live-live-situatie bij de huisartsenpatiënten en bij de patiënten van een neurosekliniek achterwege blijven. Van deze groepen was het totaal aantal geïnterviewde personen ook lager dan van de beide andere categorieën: een groter aantal proefpersonen zou een te grote belasting hebben betekend voor de desbetreffende instituten.

Bij de afname van de interviews is de volledige versie van de PSE gebruikt (9e editie). Aanvankelijk is deze volledige versie ook als zodanig in de analyses betrokken, maar uiteindelijk zijn de secties 12-15 en 18-20 conform Wing e.a. (1977) buiten beschouwing gelaten, om de volgende redenen. Op de eerste plaats bestaan de laatste items van de PSE niet uit vragen die aan de respondent worden gesteld, maar uit vragen die de interviewer door observatie moet beantwoorden en die dus via een geluidsband niet te scoren zijn.

Bovendien is zo weinig gescoord op de psychotische symptomen dat de items in sectie 12-15 niet bruikbaar bleken voor een adequate betrouwbaarheidsanalyse.

De gepresenteerde resultaten hebben derhalve betrekking op 59 items van de PSE. Slechts indien het in verband met vergelijkbaarheid met andere onderzoeken noodzakelijk is, zal in de tekst gerefereerd worden naar de uitkomsten van de analyse met alle items.

Bij de huisartsen- en neurosekliniekpatiënten werd bij elke proefpersoon 1 live-interview afgenomen en gescoord, waarna een andere beoordelaar de bandopname van dit interview nog eens scoorde. Bij de niet-patiënten en de APZ-patiënten werd het eerste interview gevolgd door een tweede interview, afgenomen door een andere interviewer. Van beide live-interviews werd een bandopname gemaakt die door een andere interviewer werd gescoord.

De interviewers – Bij het betrouwbaarheidsonderzoek waren 10 interviewers betrokken, 9 vrouwen en 1 man. Acht van hen waren afgestudeerde of bijna afgestudeerde psychologen, één was bijna arts en één was maatschappelijk werkster met Voortgezette Opleiding.

Betrouwbaarheidsstudies over de PSE hebben aangetoond dat er een

tendens bestaat bij niet in de psychiatrie ervaren interviewers, om een symptoom eerder als aanwezig te beschouwen dan medisch c.q. psychiatrisch geschoolden (Wing, 1977; Slooff, 1983). Echter, ook is aangetoond dat een grondige training en uitvoerig onderling overleg tijdens het werken met de PSE ertoe kunnen leiden dat beide groepen tot consensus komen (Cooper, 1977).

Gezien de ongelijke verdeling van disciplines, zullen we deze bevinding niet repliceren in ons onderzoek. We zijn ervan uitgegaan, dat de training dusdanig is geweest, dat verschillen, veroorzaakt door de discipline, zijn geminimaliseerd.

Vooraanvang van de training hebben de interviewers zich vertrouwd kunnen maken met de principes van de PSE en de inhoud van de vragenlijst en de handleiding. De training duurde een week en vond plaats in het Psycho-Medisch Streekcentrum Vijverdal te Maastricht onder auspiciën van Verhey die ruime ervaring heeft in de PSE-didactiek. In die periode scoorden de interviewers een tiental video-opnames en live-PSE's en bespraken hun resultaten uitvoerig met de trainer tot een communis opinio was bereikt.

Gedurende een half jaar na de training vonden regelmatig evaluaties plaats met een eveneens in de PSE techniek geschoolde psychiater. Deze bijeenkomsten hadden tot doel zowel de onderlinge consensus als de overeenstemming met de officiële PSE-lijn op peil te houden. Het onderzoek vond plaats in de periode december 1982/januari 1983. Enkele interviews vonden plaats in april 1983.

Analysemethode – Gekozen is voor een overeenstemmingsmaat die aangeeft in welke fractie van de items men het met elkaar eens is, waarbij gewichten toegekend zijn aan de verschillende vormen van niet-overeenstemming. Aan een verschil van opvatting over aan- of afwezigheid van een symptoom werd een zwaarder gewicht toegekend dan aan een verschil in opvatting over de mate waarin het kenmerk aanwezig was.

Bij patiënten is deze gewogen relatieve overeenstemming nog gecorrigeerd voor toevallige overeenstemming, dat wil zeggen de overeenstemming die gegeven de marginale scoreverdelingen te verwachten is op grond van het toeval. Dit levert dan de Kappa (Cohen, 1962), waarvan de waarden kunnen variëren van -1 (overeenstemming lager dan op basis van toeval valt te verwachten) tot +1 (volledige overeenstemming). Bij betrouwbaarheidsonderzoek rond de PSE is praktisch steeds gebruik gemaakt van Cohens Kappa als overeenstemmingsmaat.

Omdat bij niet-patiënten maar sporadisch wordt gescoord hebben de marginale verdelingen geen of weinig spreiding; correctie voor toeval heeft dan weinig zin. Zouden we ook bij niet-patiënten de Kappa berekenen dan blijkt dit te kunnen resulteren in een waarde van nul of zelfs negatieve waarde voor Kappa, ondanks het feit dat twee interviewers het in zeer vergaande mate met elkaar eens zijn over het niet

aanwezig zijn van de symptomen. Omdat Kappa in deze situatie de feitelijke overeenstemming dus slecht weergeeft, gebruiken wij daar de gewogen relatieve overeenstemming.

De resultaten

De betrouwbaarheid van de PSE op interviewniveau – Om de betrouwbaarheid tussen de interviewers te kunnen berekenen, is voor alle live-live-combinaties en voor alle live-tape-combinaties Kappa berekend, met uitzondering van de combinaties in de niet-patiënten-categorie.

Van 2 interviews zijn de bandopnames mislukt, zodat voor wat betreft de live-tape-combinaties niet overal de geplande aantallen in de tabellen staan.

Bestudering van de afzonderlijke combinaties geeft geen aanleiding te veronderstellen dat één van de interviewers een systematisch afwijkende stijl van interviewen heeft. Wel kan gezegd worden dat in de live-live-situatie (APZ-patiënten), twee combinaties van interviewers erg lage overeenstemming vertonen. In de live-tape-situatie geldt dit voor 3 combinaties, waarin steeds dezelfde interviewer betrokken was. In andere situaties en in andere combinaties kwamen de betrokken interviewers echter wel tot een aanvaardbare overeenstemming. Gedetailleerde presentatie van deze gegevens zou in het kader van dit artikel te ver voeren.

In tabel 1 is voor elke interviewer de gemiddelde Kappawaarde weergegeven, dat wil zeggen de gemiddelde Kappa van alle combinaties waarbij de desbetreffende interviewer betrokken was. Dit geldt zowel

Tabel 1: Inter-raterbetrouwbaarheid per interviewer bij patiënten.

interviewer	live-live			live-tape		
	N	K	sd	N	K	sd
1	4	0.31	0.16	11	0.43	0.33
2	4	0.52	0.13	12	0.72	0.15
3	4	0.25	0.17	11	0.55	0.33
4	4	0.50	0.41	12	0.67	0.36
5	4	0.39	0.13	11	0.62	0.14
6	4	0.29	0.04	12	0.61	0.15
7	4	0.58	0.29	12	0.68	0.23
8	4	0.40	0.30	13	0.70	0.25
9	4	0.44	0.06	12	0.58	0.15
10	4	0.19	0.08	12	0.54	0.17

N = aantal interviews waarover Kappa is berekend.

K = gemiddelde Kappawaarde.

sd = standaarddeviatie.

voor de live-live-combinaties als voor de live-tape-combinaties.

De Kappawaarden voor het live-tape-design zijn redelijk.

We zien echter dat de gemiddelde Kappa's van de live-live-interviews variëren van 0.19 tot 0.58. Kappa is gemiddeld niet erg hoog voor wat betreft het live-live-design. Gezien het feit, dat ook andere onderzoekers dergelijke resultaten melden is een hogere betrouwbaarheid in dit design kennelijk moeilijk te realiseren.

Pen c.s. (1977) hebben gesuggereerd, dat de beoordelaars na hun training minstens 25 keer de PSE moeten hebben afgenomen, daarbij voortdurend ruggespraak houdend met de andere interviewers, voordat van een redelijk betrouwbare afname sprake kan zijn.

Een andere mogelijkheid is, dat, waar andere veel gebruikte maten bij betrouwbaarheidsonderzoek een te rooskleurig beeld geven (met name de associatiematen), Kappa wellicht een wat sombere voorstelling van zaken geeft. 'Kappa does not reflect degree of agreement that the characteristic is absent', zoals Wing c.s. (1977) het uitdrukten. Bij de niet-patiënten varieerde de gevonden percentageovereenstemming van 0.82 tot 1 in de live-live-situatie en van 0.88 tot 1 in de live-tape-situatie.

Om na te gaan of er verschillen zijn tussen de verschillende patiëntencategorieën is de inter-raterbetrouwbaarheid ook per groep bekeken. We zijn uitgegaan van dezelfde Kappa's, maar de gemiddelden zijn nu niet per interviewer maar per subgroep berekend. Zoals gezegd is voor de niet-patiënten een andere maat gekozen, namelijk de gewogen relatieve overeenstemming.

De Kappa's in de live-live-situatie zijn inderdaad significant lager dan die in de live-tape-situatie (eenzijdige rangtekentoets, $p = 0.003$). Zowel op grond van de onder 'design' uiteengezette redenering als op basis van eerder onderzoek was het in hoge mate voorspelbaar dat de laatste groep een hogere betrouwbaarheid zou geven.

Zeer opvallend is de hoge betrouwbaarheid van de interviews met de

Tabel 2: Inter-raterbetrouwbaarheid voor de verschillende populaties.

	live-live			live ₁ -tape ₁			live ₂ -tape ₂		
	N	K	sd	N	K	sd	N	K	sd
huisartspatiënten		n.v.t.		10	0.59	± 0.29		n.v.t.	
APZ-patiënten	20	0.39	± 0.22	19	0.59	± 0.27	20	0.57	± 0.24
neurosepatiënten		n.v.t.		10	0.75	± 0.05		n.v.t.	
	N	O	sd	N	O	sd	N	O	sd
niet-patiënten	20	0.95	± 0.06	19	0.96	± 0.03	20	0.97	± 0.04

N = aantal interviews waarover Kappa is berekend.

K = gemiddelde Kappawaarde.

O = relatieve overeenstemming.

sd = standaarddeviatie.

neurosekliepatiënten. De overeenstemming uitgedrukt in Kappa is significant hoger bij de neurosekliepatiënten dan bij de APZ-patiënten (tweezijdige Wilcoxon-toets, p -waarde = 0.02). Alhoewel het verschil tussen huisartspatiënten en neurosekliepatiënten van dezelfde orde van grootte is, is het verschil niet significant ($p = 0.14$), mogelijk als gevolg van het kleinere aantal waarnemingen. Deze resultaten stemmen overeen met de bevinding van Slooff e.a. (1983), dat neurotische symptomen over het algemeen betrouwbaarder worden geregistreerd dan psychotische symptomen.

De betrouwbaarheid van de PSE voor de verschillende patiëntencategorieën op item niveau – We komen nu toe aan de vraag of de betrouwbaarheid van de PSE verschilt per patiëntencategorie op itemniveau.

Kappa is hier op geheel andere wijze berekend dan in de voorgaande paragraaf over inter-raterbetrouwbaarheid. Daar zijn de scores van twee complete interviews met dezelfde respondent met elkaar vergeleken, terwijl hier steeds één item als uitgangspunt is genomen. Van de hele subgroep zijn de scores op dit item per paar (interviewer-herinterviewer, respectievelijk interviewer-observator) met elkaar vergeleken, waarna het gemiddelde is genomen van al de item-Kappa's. Onderstaande Kappa's per subgroep zijn dus gemiddelden over 59 items. De relatief lage scores van de APZ-patiënten stemmen overeen met de bevinding van Pen c.s. (1977), dat 'schizofrenen' en 'andere psychosen' een significant hogere moeilijkheidsgraad hadden dan andere patiënten, dat wil zeggen, dat een hoge betrouwbaarheid van de PSE hier moeilijker te realiseren is. Als mogelijke verklaring noemt men de vaagheid of de inadequaatheid van vele door psychotici gegeven antwoorden. Omdat bij deze categorie patiënten het gebruik van alle items van de PSE (dus inclusief de secties 12-15 en 18-20) meer voor de hand zou liggen willen we hier nog aan toevoegen dat de betrouwbaarheid bij deze groep nauwelijks hoger respectievelijk lager wordt wanneer we

Tabel 3: Gemiddelde betrouwbaarheid per patiëntencategorie berekend over de items.

	live-live			live ₁ -tape ₁			live ₁ -tape ₂		
	N	K	sd	N	K	sd	N	K	sd
huisartspatiënten		n.v.t.		59	0.67	0.38		n.v.t.	
APZ-patiënten	59	0.35	0.31	59	0.57	0.32	59	0.61	0.30
neurosekliepatiënten		n.v.t.		59	0.69	0.31		n.v.t.	

N = aantal items waarover Kappa is berekend.

K = gemiddelde Kappawaarde.

sd = standaarddeviatie.

dat doen, namelijk $K = .36$ bij het live-live-design en $K = .51$ in de live-tape-situatie.

Wat verder opvalt is de relatief hoge betrouwbaarheid van de live-tape-scores van de huisartsen en neurotische patiëntencategorieën ten opzichte van de APZ-patiënten (tweezijdige rangtekentoets, $p = 0.04$ respectievelijk $p = 0.01$). Bij de neurosekliekcategorie zou dat te verklaren zijn door de duidelijke toelatingseisen die de betreffende kliniek hanteert. Dit gegeven komt ook tot uiting in de ID-scores van de neurotische respondenten (zie de volgende paragraaf) en in de hoge inter-raterbetrouwbaarheid van de betreffende populaties.

Een dergelijke verklaring gaat echter voor de huisartsenpopulatie niet op, omdat het hier juist een zeer diffuse en onduidelijke patiëntengroep betreft. Men mag tenminste aannemen dat 10 willekeurig uit een huisartsenpraktijk geselecteerde patiënten niet als een homogene groep zijn te beschouwen voor wat betreft psychiatrische symptomatologie. Misschien speelt hier een rol, dat bij deze groep een grote bereidheid bestaat om zich te uiten. Men is naar de huisarts gegaan met een bepaalde klacht, en heeft zich wellicht voorbereid op het naar voren brengen van deze klacht.

Een vergelijking van ID-scores in de verschillende populaties —

Tenslotte willen we in deze uiteenzetting over betrouwbaarheidsaspecten van de PSE ook nog een validiteitsaspect aanstippen. Aangezien wij bij vier verschillende categorieën de PSE hebben afgenomen, is het mogelijk om na te gaan of de vier groepen inderdaad van elkaar verschillen. Enerzijds zegt dit iets over de keuze van *respondenten*; zo was het aanvankelijk ook bedoeld. Men kan een dergelijke vergelijking echter ook zien als een poging de PSE op haar waarde te schatten: worden erkende psychiatrische 'gevallen' ook als zodanig door dit instrument gediagnosticeerd.

Om dit aan te tonen is voor elke respondent de Index of Definition berekend. De Index of Definition (ID), die berekend wordt op basis van aantal, aard en ernst van de symptomen, geeft aan of er voldoende informatie is voor een diagnose volgens de International Classification of Diseases (ICD). De eerste vier ID-niveaus (niveau 1: geen symptomen; niveau 2 en 3: a-specifiek neurotische symptomen; niveau 4: specifiek neurotische, c.q. sleutelsymptomen) bieden onvoldoende gegevens om een dergelijke diagnose te rechtvaardigen. Bij niveau 5 (drempelgevallen) en hoger (niveaus 6, 7 en 8) is het mogelijk de samenvatting van de symptoomscores te transformeren in ICD-termen. Hieronder volgen de ID-scores, berekend over alle items van de PSE, die de respondenten in de verschillende categorieën hebben behaald. Alleen de ID-scores van de eerste live-interviews zijn weergegeven. (De correlatiecoëfficiënt (Spearman) tussen de ID-scores bij het eerste en tweede live-interview - niet-patiënten en APZ-patiënten - bedraagt 0.73, $p < 0.001$).

niet-patiënten	11111222222222233444
huisartspatiënten	1112222266
APZ-patiënten	1222234455667777788
neurosekliepatiënten	5566777788

We zien dat binnen de niet-patiëntengroep geen enkele ID-score van 5 of hoger voorkomt. In de huisartsencategorie zijn dat er 2 van de 10 (20%), in de APZ-categorie 12 van de 20 (60%) en in de neurosekliekcategorie zijn alle scores 5 of hoger (100%). We kunnen dus aannemen dat onze vooronderstellingen met betrekking tot de verschillende categorieën juist waren. Hoewel het relatief kleine aantallen zijn, zien we een toenemende graad van diagnosticeerbaarheid naarmate het om een duidelijker populatie gaat. Opvallend is vooral de groep neurotische patiënten, die allen een ID-score van 5 of hoger hebben. Dit illustreert het gegeven, dat deze score niet zozeer de ernst van de aandoening aangeeft, als wel de mate van diagnosticeerbaarheid. Men zou kunnen stellen dat de score alleen de ernst aangeeft voor zover men deze in termen van diagnosticeerbaarheid opvat.

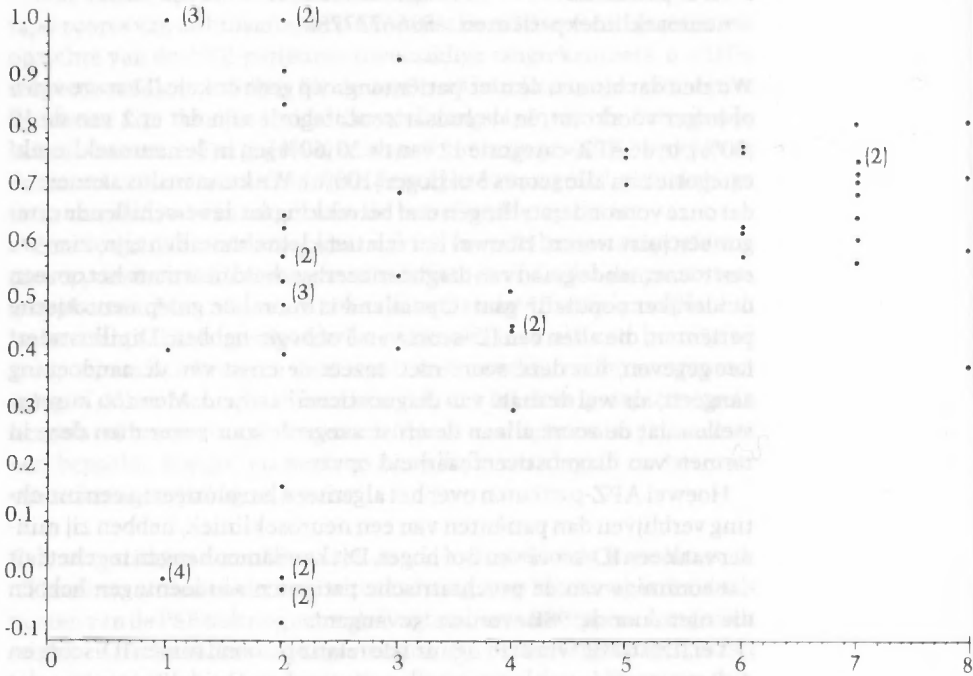
Hoewel APZ-patiënten over het algemeen langduriger in een inrichting verblijven dan patiënten van een neurosekliekcategorie, hebben zij minder vaak een ID-score van 5 of hoger. Dit kan samenhangen met het feit dat sommige van de psychiatrische patiënten aandoeningen hebben die niet door de PSE worden 'gevangen'.

Ter illustratie wordt in figuur 1 de relatie getoond tussen ID-score en de Kappawaarde van het eerste (live-)interview. Duidelijk is te zien dat Kappa bij de lage ID-scores zeer wisselende waarden aanneemt, hetgeen onze keus voor een andere overeenstemmingsmaat dan Kappa bij de niet-patiënten ondersteunt. Bij hogere ID-scores is de spreiding van Kappa kleiner. Het is echter niet zo, dat bij een hogere ID-score ook de Kappawaarde duidelijk toeneemt.

Vergelijking met de resultaten van ander onderzoek

Over het algemeen kunnen we zeggen dat onze resultaten aanvaardbaar zijn. De PSE blijkt een voldoende betrouwbaarheid op te leveren voor al de onderzochte patiëntencategorieën. Kappa, die een redelijk niveau bereikt, geeft aan dat de PSE niet alleen voldoet in klinische settings met psychiatrische patiënten, maar ook bruikbaar is bij andere patiëntengroepen. Bij niet-patiënten is Kappa als maat van overeenstemming minder bruikbaar, maar op grond van de gewogen relatieve overeenstemming kunnen we ook hier van een redelijke hoge betrouwbaarheid spreken (betrouwbaarheid betekent bij deze categorie in feite: men is het in voldoende mate eens over de afwezigheid van symptomen). Omdat voor 'betrouwbaarheid' geen absoluut minimum niveau te noemen is waaraan een instrument moet voldoen, kunnen we onze cijfers eigenlijk alleen interpreteren in het licht van andere betrouw-

Figuur 1: De relatie tussen de ID-score en de Kappa. De ID-scores van de eerste live-interviews. De Kappa's tussen de eerste live- en tape-interviews.



baarheidsonderzoeken. Het is echter niet eenvoudig een vergelijking te maken, omdat in de meeste onderzoeken analyses worden gedaan met alle items van de PSE, terwijl wij bij de analyses de secties 12-15 en 18-20 buiten beschouwing hebben gelaten.

Een tweede moeilijkheid is het feit, dat de PSE vooral is toegepast op APZ-patiënten. Dat heeft tot gevolg dat voor de niet-klinische groep, en patiënten van een neurosekliek, nauwelijks vergelijkingsmateriaal voorhanden is.

Een laatste probleem wordt gevormd door de wijze waarop Kappa is berekend in de diverse studies. We hebben al uiteengezet, dat in het onderhavige onderzoek op twee manieren de betrouwbaarheid is berekend: de inter-raterbetrouwbaarheid en de betrouwbaarheid van de PSE op item-niveau. De eerste levert geen problemen op bij vergelijking, omdat Kappa hier slechts op één manier te berekenen valt. De betrouwbaarheid van de PSE als instrument is echter over vele niveaus te berekenen. De meest elementaire is op item-niveau. Daarnaast is het ook mogelijk, Kappa te berekenen over de syndromen, over de CATEGO-klassen, over de totale score en over de ID-score.

In tabel 4 hebben we een aantal onderzoeken op een rijtje gezet die

enigszins vergelijkbaar zijn met het onderhavige. Waar mogelijk hebben we geprobeerd Kappa op item-niveau te achterhalen. In het onderzoek van Slooff c.a. zijn de Kappa's wel op item-niveau berekend, maar vervolgens in 3 groepen gegroepeerd weergegeven. We zijn zo vrij geweest hiervan het (gewogen) gemiddelde te berekenen. Waar dergelijke reconstructies niet te maken waren, is vermeld hoe Kappa was berekend in het betreffende onderzoek.

De tabel bevat informatie over de populatie, waarop de verschillende onderzoeken betrekking hebben, het design dat is toegepast (live-live of live-tape), het aantal items waarmee geanalyseerd is (lang of kort) en de manier waarop Kappa is berekend (over de interviewers, de items, de syndromen of de totaalscore).

Discussie

1. Wat betreft de inter-raterbetrouwbaarheid kunnen we zeggen dat de resultaten bevredigend zijn. De live-tape-resultaten liggen gemiddeld op .61, een aanvaardbaar niveau. Ook de Kappawaarden in de live-live-situatie zijn redelijk, in vergelijking met andere onderzoeken.

Wat verder opvalt is de hoge inter-raterbetrouwbaarheid bij de neurotische patiëntencategorie. Deze groep geeft waarschijnlijk duidelijker en meer eenduidige antwoorden dan de respondenten uit de overige categorieën.

De niet-patiëntencategorie is moeilijk vergelijkbaar met de andere categorieën, omdat we hier voor een andere overeenstemmingsmaat hebben gekozen. De gewogen relatieve overeenstemming geeft aan, dat ook deze interviews voldoende betrouwbaar worden afgenomen. Het feit, dat alle interviews zijn afgenomen en gescoord door niet-psychiaters lijkt de betrouwbaarheid niet in de weg te hebben staan.

2. De betrouwbaarheid van de PSE bij verschillende patiëntencategorieën is redelijk te noemen. We kunnen hieruit concluderen dat het gebruik van de PSE niet tot klinische psychiatrische patiënten beperkt hoeft te blijven, maar ook verantwoord is bij andere patiëntengroepen, en zelfs bij niet-patiënten.

3. De ID-scores geven een frappant beeld van de mate van diagnoseerbaarheid bij de verschillende groepen. Geheel volgens verwachting is bij de niet-patiënten geen enkele score van 5 of hoger, terwijl de neurosekliek patiënten alle in de duidelijke diagnosticeerbare categorie vallen.

Bij ID-scores van 1 t/m 3 vertonen de Kappawaarden een grote spreiding, waardoor duidelijk wordt dat bij bevolkingscategorieën waarin weinig symptomatologie te verwachten valt, een andere maat van overeenstemming meer gewenst is.

4. De hoge betrouwbaarheid van de PSE bij de neurosekliekpatiënten kan op verschillende manieren geïnterpreteerd worden. In de eerste plaats is mogelijk dat de PSE-vragen over neurotische problematiek

aanleiding geven tot een éénduidige respons. Een tweede mogelijkheid is dat de interviewers als gevolg van algemene ervaringen/of specifieke training op dat gebied eenzelfde interpretatiekader hanteren. Tenslotte is mogelijk dat de respondenten de vragen herkennen en beantwoorden als aansluitend bij hun specifieke problematiek.

Conclusie

De PSE lijkt een geschikt instrument voor onderzoek in een open populatie, die toch in meerderheid bestaat uit niet-patiënten, en waarin de voorkomende problematiek, zo leert de ervaring, meer neurotisch dan psychotisch van aard is.

Literatuur

- Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cooper, J.E., J.R.M. Copeland, G.W. Brown, T. Harris and A.J. Gourlay (1977), Further studies on interviewer training and interrater reliability of the Present State Examination (PSE). *Psychological Medicine*, 7, 517-532.
- Helzer, J.E., L.N. Robins, M. Tableson, R.A. Woodruff, Th. Reich, E.D. Wish (1977), Reliability of Psychiatric Diagnosis. I. A methodological Review. *Arch. Gen. Psychiatry*, Vol. 34, 129-133.
- Hodiamont, P.P.G. (1983), Het 'wat' in de psychiatrische epidemiologie. *Tijdschrift voor psychiatrie*, 25, 6, 391-405.
- Kendell, R.E., B. Everitt, J.E. Cooper, N. Sartorius, M.E. David (1968), The reliability of the Present State Examination. *Social Psychiatry*, 3, 123-129.
- Pen, A., W.A. Arrindell, W. Lavant, J. Pols (1977), De betrouwbaarheid van de 'Present State Examination' (een replicatiestudie). *Tijdschrift voor Psychiatrie*, 19, 375-87.
- Sanson-Fisher, R.W., C.J. Martin (1981), Standardized Interviews in Psychiatry: Issues of Reliability. *British Journal of Psychiatry*, 139, 138-143.
- Schouten, H.J.A. (1982), Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica*, 36, nr. 2, 45-62.
- Slooff, C.J., W.R.E.H. Mulder-Hajonides van der Meulen, R.H. van den Hoofd-akker (1983), De Nederlandse vertaling van de Present State Examination. Deel I. Betrouwbaarheidsaspecten. *Tijdschrift voor Psychiatrie*, 25, 1983/3. Deel II. De relatie tussen CATEGO-klassen en categorieën van de 'International Classification of Diseases'. *Tijdschrift voor Psychiatrie*, 25, 1983/4.
- Spitzer, R.L. and J.L. Fleiss (1974), A reanalysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry*, 125, 341-7.
- Tax, B., König-Zahn, P.H.J.M. Heydendaal, L.H. Boerma, J.W. Furer, E.J.H. ter Heine, P.P.G. Hodiamont, J.M.G. Persoon, S.H.J. Veling, *Regioprospect Nijmegen*. Een longitudinaal onderzoek in de bevolking naar ziekte- en hulpzoekgedrag bij somatische klachten, psychiatrische symptomen en psychosociale problemen. *Gezondheid en Samenleving*, 1, 38-45.
- Verhey, F.H.M. (1981), *Langzame hersenpotentialen en gemeten gedrag*. Aca-

- demisch proefschrift, Utrecht.
- Wing, J.K., J.L.T. Birley, J.E. Cooper, P. Graham, A.D. Isaacs (1967), Reliability of a Procedure for Measuring and Classifying 'Present Psychiatric State'. *British Journal of Psychiatry*, 113, 499-515.
- Wing, J.K., J.M. Nixon, S.A. Mann and J.P. Leff (1977), Reliability of the PSE (ninth edition) used in a population study. *Psychological Medicine*, 7, 505-516.
- Wing, J.K., S.A. Mann, J.P. Leff and J.M. Nixon (1978), The concept of a 'case' in psychiatric population surveys. *Psychological Medicine*, 8, p. 203-217.
- Wing, J.K. (1980), Standardizing clinical diagnostic judgements. The PSE-CATEGO system. *Australian and New Zealand Journal of Psychiatry*, 14, 17-20.

De auteurs zijn dank verschuldigd aan: drs. S.H.J. Veling en drs. N. Sijben voor hun steun bij de opzet en het verloop van het onderzoek; dr. F. Verhey voor de training van de interviewers; de interviewers zelf, voor hun inzet tijdens het veldwerk; P. Vreuls voor de programmering; Gezondheidscentrum Lindenholt te Nijmegen, Huize Padua te Boekel, Psychiatrisch Ziekenhuis Reinier van Arkel te 's-Hertogenbosch en kliniek Overwaal te Lent voor hun medewerking. Tenslotte de respondenten zelf voor hun bereidheid mee te doen aan de interviews.