

De betrouwbaarheid van de 'Present State Examination' (een replicatiestudie)

door A. Pen, W. A. Arrindell, W. Lawant en J. Pols

Inleiding

De geringe tussenbeoordelaarsbetrouwbaarheid van diagnostici bij het beoordelen van psychiatrische toestandsbeelden is in de afgelopen twee decennia onderwerp geweest van veel discussie en onderzoek. Overzichten van de literatuur over de betrouwbaarheid van psychiatrische diagnostiek vindt men o.a. bij Ash (1949), Beck (1962), Kreitman (1961), Rooymans (1969) en Spitzer (1967). Bij pogingen om de betrouwbaarheid van bestaande classificatiesystemen (ICD, DSM) te verbeteren stuit men steeds op een fundamenteel probleem:

De klassegrenzen van de gebruikelijke ziekteklassen worden gedefinieerd met behulp van een aantal variabelen van zeer verschillende aard zoals: symptoomconstellaties, verloop in de tijd, genetische variabelen etc. De samenhangen tussen deze variabelen zijn echter slechts zeer ten dele empirisch bevestigd. Het is daarom de vraag of de traditionele ziekteklassen een weerspiegeling zijn van bestaande patiëntengroepen. Of anders geformuleerd: het valt te betwijfelen of men met behulp van de gebruikelijke classificatiesystemen in staat is 'to cut nature at the joints'. (Kendell, 1975).

Dat deze twijfel gerechtvaardigd is blijkt uit de diagnostische praktijk. Veel patiënten vallen binnen meerdere diagnostische categorieën omdat deze elkaar niet wederzijds uitsluiten. Het hier gesignaleerde probleem heeft primair te maken met het validiteitsaspect van de diagnostiek doch het heeft ook rechtsstreekse consequenties voor de betrouwbaarheid.

Bij het diagnosticeren van patiënten die bij meerdere diagnostische klassen kunnen worden ondergebracht, zal de beslissing worden bepaald door de persoonlijke bias van de onderzoeker of door het toeval.

De hier in het kort geschetste problemen hebben verschillende on-

Schrijvers zijn resp.: psychiater, hoofdmedewerker Psych. Klin. R.U. Groningen; psycholoog, R.U. Groningen; psychiater, Psych. Zhs. 'Licht en Kracht', Assen; psychiater, Psych. Zhs. 'Licht en Kracht', Assen.

derzoekers er toe gebracht om het concept van de ziekte-klasse voorlopig terzijde te leggen en zich te beperken tot een beschrijvende diagnostiek van symptomen. Deze gang van zaken heeft geleid tot de ontwikkeling van twee typen diagnostische instrumenten, de psychiatrische beoordelingsschalen en de gestructureerde psychiatrische interviews.

Psychiatrische beoordelingsschalen

Een bespreking van de verschillende thans in omloop zijnde schaaltypen valt buiten het bestek van dit artikel, een overzicht van de belangrijkste en best onderzochte schalen vindt men bij Lysterly (1973). Met behulp van beoordelingsschalen kunnen de kwalitatieve en de kwantitatieve aspecten van de meest voorkomende psychopathologische verschijnselen enigszins worden geobjectiveerd waardoor de variantie ten gevolge van een persoonlijke beoordelaarsbias kan worden gereduceerd. De meeste diagnostische schalen definiëren gedrag op symptoom-niveau. Verschillende auteurs (Lorr, 1973; Overall en Klett, 1972) hebben echter procedures ontwikkeld om de meest voorkomende symptomcombinaties binnen patiëntenpopulaties op te sporen, zodat individuele patiënten kunnen worden beschreven in termen van empirische syndromen. Een nadeel van beoordelingsschalen is de betrekkelijke ongestructureerdheid van het proces van informatiewinning – het interview en de observatie – zodat een belangrijke bron van meningsverschillen tussen beoordelaars aanwezig blijft.

Aan dit bezwaar wordt tegemoet gekomen door een tweede type van diagnostische instrumenten, het gestructureerde psychiatrische interview.

Gestructureerde psychiatrische interviews

Met behulp van deze instrumenten kan een van te voren gedefinieerd veld van psychopathologie worden geëxploreerd; doordat de vragen zijn gestructureerd is de wijze van informatiewinning vastgelegd, zodat de invloed van een aantal interviewvariabelen zoals interviewstijl, theoretische achtergronden, etc., tot een minimum beperkt blijft.

De Present State Examination

Het tot nu toe best beproefde gestandaardiseerde interview is de Present State Examination (PSE). De PSE werd gedurende afgelopen 12 jaar ontwikkeld in de Medical Research Council van het Institute of Psychiatry te Londen (Wing e.a. 1967). De vroegste versies werden ontworpen voor de beoordeling van patiënten die leden aan functionele psychosen: affectieve psychosen en schizofrenieën. Aan de hand van een aantal betrouwbaarheids- en validiteitsstudies werden voortdurend wijzigingen aangebracht, zodat thans de negende versie van de schaal ter beschikking staat. Deze bestaat uit 140 items, waarin de meest voorkomende psycho-pa-

thologische verschijnselen aan de orde komen. De eerste 105 items zijn zgn. vraag-items, waarin gestandaardiseerde vragen naar subjectieve symptomen zijn geformuleerd. Dikwijls bevatten de items alternatieve vragen, die gesteld kunnen worden wanneer daartoe aanleiding bestaat, bijvoorbeeld wanneer de interviewer niet zeker is of de patiënt de vraag begrepen heeft. Bovendien heeft de interviewer de vrijheid om indien dit strikt noodzakelijk is zelf aanvullende vragen te bedenken. De laatste 25 items hebben betrekking op afwijkend gedrag. Alle items kunnen worden beoordeeld op een 3-puntsschaal (0 = afwezig, 1 = matig ernstig, 2 = ernstig). Indien een antwoord niet duidelijk is wordt gescoord met een 8.

Het interview bestrijkt een tijdsbestek van 4 weken voorafgaande aan het moment van afname. De 140 items zijn arbitrair ingedeeld in 20 secties die elk gericht zijn op een afzonderlijk gebied van psychopathologie. Voor elke sectie kan door optelling van de itemscores een sectiescore worden berekend. Elke sectie wordt ingeleid met een aantal oriënterende vragen. Indien de antwoorden op deze vragen negatief luiden, kan met een grote mate van zekerheid worden aangenomen dat in het betreffende symptoomgebied geen verdere afwijkingen aanwezig zijn, zodat de resterende vragen achterwege kunnen blijven. Indien het laatste het geval is wordt gebruik gemaakt van een afsluitpunt. Als het afsluitpunt is gebruikt krijgen alle resterende items, binnen de sectie, die samen de zgn. afsluitsectie vormen, een score 0. Op deze wijze kan het interview in een tijd van 3 kwartier tot 1½ uur worden afgenomen.

De eerste 10 secties bevatten vragen over verschijnselen die zich vooral voordoen bij stoornissen van neurotische aard, bijv. tobben, diverse manifestaties van angst, fobieën, subjectieve denk- en concentratiestoornissen, depressieve stemming, insufficiëntie- en schuldgevoelens, prikkelbaarheid, dwangsymptomen, etc. Daarop volgen 2 secties waarin gevraagd wordt naar verschijnselen die dikwijls bij meer ernstige decompensatietoestanden zoals depersonalisatie, waarnemingsstoornissen van niet-hallucinatoire aard, betrekkingsideeën, etc. In de secties 13 t/m 17 komen symptomen van hoofdzakelijk psychotische aard aan de orde. Allereerst een aantal subjectieve denkstoornissen zoals het gevoel van gedachtenonttrekking, gedachtenuitbreiding, beïnvloedingsideeën en andere verschijnselen, die door Kurt Schneider tot de 'symptomen van de eerste orde' worden gerekend. Vervolgens wordt gevraagd naar hallucinaties van verschillende aard en tenslotte naar wanen. Daarmee wordt het interviewgedeelte van de schaal afgesloten.

De laatste 2 secties worden gevuld door 20 items waarin het gedrag en de verschillende kwaliteiten van het affect tijdens het interview worden beoordeeld.

Het scoreformulier van de PSE bevat een groot aantal definitieve regels, aanwijzingen voor het scoren van items en suggesties voor alternatieve vragen.

Hoewel de PSE ontworpen is om het psychiatrisch interview te structureren, wordt aan de interviewer een zekere mate van vrij-

heid gegeven om eigen formuleringen te kiezen indien de toestand van de patiënt daartoe aanleiding geeft. De auteurs benadrukken het feit dat de PSE geen 'inquiry' is doch een klinisch-diagnostische schaal. Dit houdt in dat in laatste instantie de onderzoeker beoordeelt of een symptoom al of niet aanwezig is. In een uiterste geval kan het oordeel van de onderzoeker zelfs tegengesteld zijn aan de mededelingen die de patiënt doet aan de hand van de gestelde vragen. Wanneer bijvoorbeeld het bestaan van hallucinaties wordt ontkend doch uit het gedrag van de patiënt duidelijk is af te leiden dat hij stemmen hoort, zullen hallucinaties positief worden beoordeeld.

Het klinische karakter van de PSE blijkt ook uit de opbouw van de verschillende secties. Symptomen die minder ernstige stoornissen van bepaalde symptoomgebieden benoemen gaan vooraf aan die welke ernstiger of meer uitgebreide stoornissen beschrijven. Zo wordt in het gedeelte dat over wanen gaat eerst aandacht geschonken aan de verschillende aspecten van de waanstemming, waarna items volgen waarin de mate van systematisering van de waan wordt beoordeeld. Afzonderlijke secties zijn gereserveerd voor het noteren van gegevens over het korte inleidende gesprek dat aan het interview voorafgaat en voor het waarderen van de kwaliteit van het interview.

De 7e versie van de schaal werd in negen talen vertaald en in twee internationale studies getoetst op een aantal aspecten van betrouwbaarheid en validiteit (Cooper e.a. 1972; Lin e.a. 1973). De meest recente 9e versie is voorzien van een uitvoerige glossary en de beschrijving van een computerprogramma (CATEGO), met behulp waarvan het score-materiaal kan worden geïnterpreteerd in termen van experimentele syndromen (Wing e.a. 1974).

Replicatie van een betrouwbaarheidsstudie

Door Kendell e.a. (1968) werd de betrouwbaarheid van de PSE onderzocht. Hij mat de tussenbeoordelaars-overeenstemming van onafhankelijke beoordelaars die een willekeurige steekproef van patiënten met behulp van de PSE onderzochten. De resultaten waren bevredigend, doch daarbij dient te worden opgemerkt dat de beoordelaars langdurig waren getraind en hun observaties verrichtten binnen een research-setting. De toepassing van de schaal is echter dermate ingewikkeld dat de vraag rees of de PSE een geschikt routine-instrument is dat kan worden gebruikt door minder intensief getrainde beoordelaars binnen een normale klinische setting. Om deze vraag te kunnen beantwoorden werd Kendell's onderzoek door ons gerepliceerd.

Daarbij was de gevolgde procedure identiek aan die van de oorspronkelijke studie, afgezien van het feit dat de beoordelaars slechts summier waren getraind, en hun beoordelingen maakten in het kader van hun normale klinische werkzaamheden. Het onderzoek vond enige tijd geleden plaats, zodat niet de 9e, doch de 8e

versie van de schaal werd gebruikt (terwille van de actualiteit hebben wij ons bij de beschrijving gehouden aan de 9e versie).

De structuur van de 8e versie is geheel gelijk aan die van de 9e. Hij is echter belangrijk uitgebreider (484 items verdeeld over 27 secties) omdat in de 9e versie een groot aantal items welke voornamelijk betrekking hebben op inhoudelijke details van wanen, zijn weggelaten.

De grootte van onze patiënten-steekproef was iets kleiner dan die van Kendell. Zijn groep bevatte 37 patiënten waarvan er 25 werden geherinterviewd. Onze steekproef bevatte 28 patiënten die allen werden geherinterviewd.

Methode

De 8e versie van de PSE werd met toestemming van de auteur door ons uit het Engels in het Nederlands vertaald.

De vertaling werd gecorrigeerd door een klinisch geschoolde vertaler die twee-talig was. Drie psychiaters die destijds allen werkzaam waren in het psychiatrisch centrum 'Licht en Kracht' te Assen trainden elkaar in het gebruik van de schaal. Eén van hen had een aantal trainingssessies gevolgd in de Medical Research Unit van het Institute of Psychiatry te Londen. Tijdens deze trainingsperiode werden tien patiënten gemeenschappelijk onderzocht, de scores werden besproken en getracht werd om zoveel mogelijk oorzaken van meningsverschillen op te sporen.

Vervolgens werden 28 nieuw opgenomen patiënten door de drie psychiaters onafhankelijk van elkaar onderzocht. De toewijzing van een nieuw opgenomen patiënt aan de onderzoeksgroep hing af van de vraag of op de dag waarop het interview moest plaatsvinden twee psychiaters tegelijkertijd ter beschikking waren.

De selectie was daardoor niet willekeurig, doch wel onafhankelijk van het toestandsbeeld van de patiënt. Elke patiënt werd binnen 48 uur na opname geïnterviewd door één van de psychiaters (A) terwijl een tweede psychiater (B) het interview bijwoonde en de patiënt onafhankelijk mee beoordeelde. Binnen vijf dagen na het eerste interview werd het onderzoek herhaald door een derde psychiater (C).

De drie diagnostische rollen: interviewer (A), observator (B) en herinterviewer (C) werden gelijk over de drie psychiaters verdeeld, zodat iedere psychiater elk van de drie rollen ongeveer even vaak vervulde. Het onderzoek vond plaats in een tijdsbestek van twee maal twee maanden.

De uitboek-diagnoses (ICD) van de 28 patiënten luiden: schizofrenie (N = 6), affectieve psychosen (N = 5), overige psychosen (N = 7), dementie (N = 1), neurose (N = 2), alcoholisme (N = 3), persoonlijkheidsstoornis (N = 2), niet-psychotische stoornis bij somatisch lijden (N = 2).

De opzet van de studie was zodanig, dat per patiënt steeds twee beoordelaars-combinaties konden worden vergeleken: de combinatie interviewer-observator en de combinatie interviewer-herinter-

viewer. Op voorhand mag worden verondersteld dat de overeenstemming van de eerste combinatie (interviewer vs. observator) groter zal zijn dan die van de tweede combinatie (interviewer vs. herinterviewer).

De interviewer en observator baseren hun oordeel op dezelfde steekproef van gedrag, en met name in die gevallen waar wordt geoordeeld op grond van eenduidige bevestigende of ontkennende antwoorden zijn meningsverschillen vrijwel uitgesloten. Bovendien kan er sprake zijn van een bias van de observator in de richting van de interviewer. Wanneer de interviewer bepaalde secties snel behandelt, weinig alternatieve vragen stelt of afsluitpunten gebruikt, kan de observator daaruit zijn conclusies trekken.

De interviewer-herinterviewer situatie bevat echter een aantal bronnen van discordantie die geen rol spelen bij de interviewer-observator situatie. Het toestandsbeeld van de patiënt kan gedurende het tijdsinterval tussen het eerste en het tweede interview zijn veranderd en het gedrag van de patiënt wordt mogelijk beïnvloed door herhaling van hetzelfde interview binnen enkele dagen.

Hoewel de gang van het interview bij de PSE in hoofdzaak voorgestructureerd is, werd ruimte overgelaten voor individuele variaties. Het is dan ook mogelijk dat de interviewstijl van de tweede interviewer niet geheel dezelfde is als die van de eerste interviewer, waardoor de patiënt zich tijdens het tweede interview anders kan gedragen dan tijdens het eerste.

We mogen dan ook veronderstellen dat de interviewer-observator overeenstemming een te gunstig, en de interviewer-herinterviewer overeenstemming een te ongunstig beeld geeft van de werkelijke betrouwbaarheid van de PSE.

Bij de analyse van de overeenstemming tussen de verschillende beoordelaarscombinaties is gebruik gemaakt van een drietal overeenstemmingsmaten.

1 – De 'discordantiecoëfficiënt' welke wordt berekend door het aantal meningsverschillen tussen twee beoordelaars per patiënt te delen door het gemiddelde aantal pathologiescores van de beoordelaars per patiënt. Door de meningsverschillen te relateren aan het gemiddelde aantal pathologiescores wordt voorkomen dat de overeenstemming voor patiënten met weinig pathologie een geflatteerd beeld geeft. Om onderscheid te maken tussen geringe en ernstige meningsverschillen tussen beoordelaars werden steeds 2 discordantiecoëfficiënten berekend. Een 'algehele discordantie'-coëfficiënt en een 'ernstige discordantie'-coëfficiënt.

In de 'algehele discordantie'-coëfficiënt werden alle voorkomende meningsverschillen (0 vs. 1, 2, 8; 1 vs. 2, 8; 2 vs. 8) verdisconteerd terwijl bij de 'ernstige discordantie'-coëfficiënt alleen de meest belangrijke verschillen (0 en 8 vs. 1 en 2) werden betrokken.

2 – De produkt-momentcorrelaties van de somscores per sectie (sectiescores).

3 – De overeenstemmingscoëfficiënt Kappa per item (Cohen 1968). Kappa is een overeenstemmingscoëfficiënt voor nominale schalen,

die twee voordelen biedt boven de gebruikelijke overeenstemmingsmaten zoals produkt-momentcorrelaties of de percentage-overeenstemming. In de eerste plaats is Kappa gecorrigeerd voor overeenstemming welke het gevolg is van toeval. Omdat in het algemeen per patiënt slechts een relatief klein aantal items positief wordt gescoord, is het scoremateriaal van pathologieschalen in het algemeen scheef verdeeld, met als gevolg dat een belangrijk deel van de gevonden overeenstemming op toeval kan berusten. Daardoor geven de produkt-momentcorrelaties en de percentage-overeenstemming dikwijls een geflatteerd beeld. Bij Kappa is het effect van overeenstemming door toeval echter geëlimineerd. Een ander voordeel van Kappa is de mogelijkheid om door middel van het toekennen van gewichten aan scoreverschillen de omvang van de meningsverschillen in de overeenstemmingsmaat tot uitdrukking te brengen (Cohen, 1968).

De toekenning van de gewichten is arbitrair en vond op dezelfde manier plaats als in Kendell's studie: het ernstigste meningsverschil: 0 vs. 1 of 2 kreeg gewicht 3, het minder ernstige meningsverschil 1 of 2 vs. 8 kreeg gewicht 2, en de betreffende triviale verschillen 0 vs. 8 of 1 vs. 2 kregen gewicht 1. De waarden van Kappa variëren van negatief (overeenstemming slechter dan toeval) via 0 (overeenstemming gelijk aan toeval) tot 1.00 (overeenstemming volledig).

Elke score is de resultante van een drietal interveniërende variabelen: de patiënt, de aard van het item, en de interviewer. De onderzoekresultaten zullen rond deze drie variabelen worden gegroepeerd. De door Kendell gevonden overeenstemmingswaarden zullen steeds tussen haakjes worden toegevoegd.

Resultaten

1 - De patiënten. De variatie-breedte van de discordantiecoëfficiënten van de verschillende patiënten is groot. Zo varieerden de 'algehele discordantie'-coëfficiënten voor de interviewer-observator-combinaties van .20 - 1.34 (.20 - 1.10).

Bij Kendell's onderzoek bleek de discordantie voor 'schizofrene' patiënten hoger te liggen dan voor de overige patiënten ($p < 0.001$). Bij de door ons uitgevoerde toetsing werden de patiënten met de diagnose 'schizofrenie' en 'andere psychosen' (een categorie die in Kendell's materiaal niet voorkwam) samengevoegd tot één groep ($N = 13$).

De discordantiecoëfficiënten van de patiënten in deze gecombineerde groep bleken, evenals het geval was bij Kendell's groep 'schizofrenen', hoger te zijn dan die voor de overige 15 patiënten (Man-Whitney U-toets eenzijdig, $p < 0.05$). Dit verschil gold alleen voor de interviewer-observator combinaties. De discordantie voor de groep van de affectieve psychosen ($N = 5$) bleek niet significant te verschillen van die van de niet-psychotische patiënten. Deze bevinding stemt overeen met Kendell's onderzoek.

2 – De items. De Kappa-waarden per item werden berekend over alle 28 patiënten. De gemiddelde waarde van Kappa over alle items bedroeg .59 (.73) voor de interviewer-observator combinaties en .30 (.41), dus aanzienlijk lager, voor de interviewer-herinterviewer combinaties. De 484 afzonderlijke items varieerden onderling sterk in Kappa-waarde, van lager dan .00 tot maximaal 1.00.

Voor wat betreft de interviewer-observator combinaties geven de items die het gedrag van de patiënt beschrijven lagere Kappa-waarden dan de items waarin een vraag gesteld wordt.

De gemiddelde Kappa-waarde voor de gedragsitems bedroeg .34 (.49) tegenover .68 (.76) voor de vraag-items; een verschil in gemiddelde Kappa-waarde dat statistisch in hoge mate significant is ($t = 4.8, p < 0.0001$).

Dit verschil verdwijnt bij de interviewer-herinterviewer combinaties. Hier bedraagt de gemiddelde Kappa-waarde voor gedrags-items .33 (.37) en voor antwoorditems .29 (.41), een verschil dat statistisch niet significant is. Ook deze toetsingsresultaten stemmen geheel overeen met die van Kendell's studie.

Zoals bij de beschrijving van de schaal werd uiteengezet zijn de 484 items gegroepeerd in 25 secties die elk een bepaald symptoom-gebied bestrijken. Per sectie werden somscores berekend en deze werden zowel voor de interviewer-observator combinaties als voor de interviewer-herinterviewer combinaties gecorreleerd. De resultaten zijn weergegeven in tabel 1.

3 – Het gebruik van afsluitpunten. Zoals reeds werd uiteengezet kan het interview in belangrijke mate worden bekort door het gebruik van afsluitpunten. Een afsluitpunt wordt gebruikt indien de oriënterende vragen van een sectie negatief worden beantwoord, waarna de resterende vragen die samen de 'afsluit-sectie' vormen, achterwege mogen blijven en als 0 worden gescoord. De interviewer-herinterviewer situatie biedt de gelegenheid om de betrouwbaarheid van deze procedure te toetsen.

De PSE bevat 27 afsluitpunten met even veel afsluitsecties. De drie beoordelaars gebruikten gemiddeld 14 afsluitpunten zodat zij gemiddeld 13 afsluitsecties invulden. In gemiddeld 10 van de 13 gebruikte afsluitsecties werden positieve scores gemaakt. De interviewer en de herinterviewer gebruikten in 67% (80%) van de gevallen dezelfde afsluitsecties, maar ze verschilden in het gebruik van de resterende 33%.

Het blijkt dat de beslissing om afsluitpunten te gebruiken de overeenstemmingsresultaten niet nadelig beïnvloedt. De gemiddelde Kappa-waarden voor de oriënterende items, die altijd gevraagd werden, bedroeg .26 (.45). Die voor de items in de afsluitsecties .32 (.40).

Het verschil was niet significant.

4 – De beoordelaars. De verschillen in discordantie tussen de diverse beoordelaars combinaties werden vergeleken met behulp van de discordantiecoëfficiënten. De gemiddelde 'algehele discordantie'-coëfficiënt bedroeg voor de interviewer-observator combinaties

De betrouwbaarheid van de 'Present State Examination'

Tabel 1: De betrouwbaarheid van de sectie scores (prod.-mom. correlaties)

	interv. vs observator		interv. vs herinterviewer	
tobben	0.818	(0.83)*	0.645	(0.73)
spierspanning	0.875	(0.93)	0.407	(0.67)
vrij floterende angst	0.855	(0.93)	0.672	(0.59)
situatieve angst	0.872	(0.58)	0.300	(0.67)
vermijding	0.847	(0.81)	0.102	(0.65)
autonome symptomen	0.939	(0.97)	0.769	(0.83)
langz. en verward denken	0.513	(0.80)	0.340	(0.57)
vertraagde motoriek	0.775	(0.85)	0.625	(0.64)
betrekkingsideeën	0.596	(0.86)	0.581	(0.31)
zelfgevoel	0.822	(0.88)	0.416	(0.58)
depressieve stemming	0.915	(0.85)	0.614	(0.71)
somatische symptomen	0.819	(0.85)	0.707	(0.65)
prikkelbaarheid	0.522	(0.92)	0.514	(0.44)
elevation of mood, overactiviteit	0.949	(0.85)	0.724	(-0.05)
obsessies	0.946	(0.93)	0.817	(0.78)
afname interesse	0.604	(0.86)	0.058	(0.78)
concentratie	0.827	(0.94)	0.366	(0.64)
depersonalisatie	0.974	(0.96)	0.504	(0.04)
waarnemingsstoornissen	0.860	(0.81)	0.294	(-0.02)
vreemd gedrag				
katatoon gedrag	0.687	(0.72)	0.612	(0.73)
zelfverwaarlozing				
overactiviteit	0.968	—	0.757	—
traagheid van bewegingen	0.642	—	0.404	—
incoherentie van de spraak	0.567	—	0.780	—
depressieve wanen en hallucinaties	0.945	—	0.739	—

* Voor de laatste 4 secties werden door Kendell geen resultaten vermeld.

.88 (.52) en voor de interviewer-herinterviewer combinaties 1.38 (1.03).

De waarden voor de 'ernstige discordantie'-coëfficiënten bedroegen respectievelijk .52 (.35) en .99 (.80). De discordantie tussen de beoordelaars was dus voor de interviewer-observator combinaties ca. 50%, en voor de interviewer-herinterviewer combinaties ca. 30% hoger dan bij Kendell's onderzoek.

In tabel 2 zijn de discordantiecoëfficiënten voor de verschillende interviewer-observator combinaties weergegeven. Daarbij werd de vraag, wie interviewer en wie observator was, buiten beschouwing gelaten.

Het blijkt dat de overeenstemming tussen het beoordelaarspaar A vs. C groter is dan tussen de beoordelaarsparen A vs. B en B vs. C, doch het verschil is niet significant op 5% niveau. (Kruskal-Wallis, 2-zijdig). Vervolgens werd nagegaan of de identiteit van de interviewer van invloed was op de overeenstemming tussen interviewer en observator. Daarbij werd de identiteit van de observator

Tabel 2

beoordelaars-comb.	aantal interviews	gemiddelde 'algehele discordantie'-coëff.	gemiddelde 'ernstige' discordantie'-coëff.
A vs. B	9	0.91	0.55
B vs. C	10	0.92	0.56
A vs. C	9	0.80	0.44

Tabel 3

interviewer	aantal interviews	gemiddelde 'algehele discordantie'-coëff.	gemiddelde 'ernstige' discordantie'-coëff.
A (vs. B of C)	11	0.78	.46
B (vs. A of C)	8	1.10	.74
C (vs. A of B)	9	0.80	.40

buiten beschouwing gelaten. De resultaten zijn weergegeven in tabel 3.

Het blijkt dat de discordantie tussen interviewer B en de observatoren groter was dan tussen de interviewers A en C en hun resp. observatoren. Het verschil was echter wederom niet significant (Kruskall-Wallis, 2-zijdig).

5 – Systematische discordantie. Een gedeelte van de discordantie tussen beoordelaars kan het gevolg zijn van het feit dat sommige beoordelaars een lagere tolerantie-drempel hebben voor pathologie, zodat zij vaker en hoger scoren dan anderen. Bij het kwantificeren van afwijkingen kiezen zij lagere ankerpunten, waardoor systematische verschillen ontstaan die in onderlinge discussies betrekkelijk gemakkelijk zijn te elimineren. Het is daarom zinvol om bronnen van systematische discordantie op te sporen.

Voor het vergelijken van de beoordelingsdrempels van de drie beoordelaars werden de volgende indices berekend:

- het gemiddelde aantal positieve beoordelingen per beoordelaar per interview,
- het gemiddelde van de sectie-scores per beoordelaar per interview.

Bij vergelijking van het gemiddelde aantal positieve beoordelingen per interview blijkt dat beoordelaar A meer positieve beoordelingen geeft dan de interviewers B en C (Friedman, 2-zijdig $p = 0.02$). Bij vergelijking van de gemiddelde sectie scores blijkt dat deze lager zijn voor interviewer B dan voor de interviewers A en C (Friedman 2-zijdig $p < 0.0001$).

Interviewer A scoort dus vaker afwijkingen dan interviewer B en C, terwijl interviewer B gemiddeld lager scoort dan de interviewers A en C.

Discussie

Kendell's studie van de betrouwbaarheid van de 8e versie van de

PSE werd verricht onder optimale research-omstandigheden door een aantal grondig getrainde onderzoekers. Het doel van ons onderzoek was het beantwoorden van de vraag of beoordelaars met een summiere training en werkend onder minder ideale omstandigheden tot vergelijkbare resultaten konden komen. Alvorens nader op deze vraag in te gaan een aantal algemene opmerkingen:

In de loop van het onderzoek bleek dat het ook voor relatief slecht getrainde beoordelaars mogelijk is om het gestructureerde interview af te nemen op een manier die niet storend is voor de patiënt en die een goed contact tussen patiënt en beoordelaar toelaat. Een probleem voor de beginner is echter het aanvankelijke gebrek aan flexibiliteit.

Een geroutineerde onderzoeker kan zich aanpassen bij de volgorde waarin de patiënt zijn klachten brengt omdat de items in elke gewenste volgorde mogen worden behandeld. Wanneer men deze routine mist tracht men de patiënt echter te dwingen in de volgorde van de schaal waardoor het gesprek een geforceerd karakter krijgt. Ook duurt het geruime tijd voordat een beoordelaar zodanig thuis is in het gecompliceerde systeem van definities, toelichtingen en score-aanwijzingen, dat hij zijn volle aandacht bij de patiënt kan houden. De auteurs van de schaal adviseren aspirant gebruikers dan ook een trainingsperiode waarbij 20 patiënten onder supervisie worden beoordeeld. De door ons georganiseerde intervisie-training waarbij 10 patiënten werden beoordeeld bleek dan ook onvoldoende te zijn, wat tot uiting kwam in de relatief lage betrouwbaarheidsresultaten. Bij vergelijking van de door ons bereikte overeenstemming met die van Kendell's onderzoek blijkt dat onze resultaten in het algemeen slechter waren. Dit geldt zowel voor de discordantie-coëfficiënten, de Kappa-waarden als voor de correlatie-coëfficiënten.

Afgezien van de geringere onderlinge overeenstemming waren een aantal bevindingen bij ons onderzoek echter identiek aan die van Kendell. De discordantie-coëfficiënten per patiënt vertoonden ook bij ons onderzoek een grote spreiding, de 'algehele discordantie'-coëfficiënten varieerden voor de interviewer-observator combinaties van .20 tot 1.34. Er zijn dus relatief moeilijk en relatief gemakkelijk beoordeelbare patiënten.

De moeilijkheidsgraad van de groep 'schizofrenen' en 'andere psychosen' (54% van het totaal) bleek significant hoger te liggen dan die voor de overige patiënten. Waarschijnlijk is het de vaagheid of de inadequaatheid van vele door psychotici gegeven antwoorden die beoordeling zo moeilijk maken. Daarbij komt dat de gedrags-items waarvoor de overeenstemming lager ligt dan voor antwoord-items, hoofdzakelijk betrekking hebben op psychotisch gedrag. Het relatief grote aantal psychotische patiënten in onze groep heeft mogelijk een rol gespeeld bij onze relatief lage overeenstemming in vergelijking met Kendell's onderzoek.

Zijn patiëntengroep bevatte 23% 'schizofrenen' waarvoor de overeenstemming significant lager lag dan voor de overige sub-groepen.

Onze groep bevatte 54% 'schizofrenen' en 'andere psychosen' (een categorie die ontbrak in Kendell's populatie) waarvoor de overeenstemming significant lager was dan voor de overige patiënten. Het is dan ook niet uitgesloten dat onze groep een groter aantal moeilijk beoordeelbare patiënten bevatte. Onze ervaring is dat de PSE geen geschikt instrument is om ernstig verwarde patiënten te onderzoeken. Bij de beoordeling is men te zeer afhankelijk van het opvattingsvermogen van de patiënt en de begrijpelijkheid van zijn antwoorden.

Volgens Kendell kwamen dit soort ernstig verwarde patiënten bij zijn onderzoek niet voor, wat mogelijk verband houdt met een verschil in opnamestrategie tussen de instituten waarin de beide onderzoeken werden verricht.

De gemiddelde Kappa-waarden bij ons onderzoek lagen duidelijk lager dan de door Kendell opgegeven waarden (.59 vs. .73).

De hoge betrouwbaarheid van vraagitems in vergelijking met observatie-items in de interviewer-observatie combinaties, is conform de bevindingen van Kendell. Voor het verdwijnen van dit verschil bij de interviewer-herinterviewer situatie geldt hetzelfde. Kendell veronderstelt dat dit effect het gevolg is van het feit dat patiënten tijdens het herinterview anders antwoorden dan tijdens het eerste interview, terwijl hun gedrag relatief constant blijft. Een andere verklaring kan echter zijn dat de observator met name voor de vraagitems gebiased is in de richting van de interviewer. Met name wanneer de interviewer een afsluitpunt gebruikt weet de observator dat de interviewer alle volgende items van de sectie met 0 beoordeelt.

De door ons gevonden correlaties van de sectie-scores benaderden Kendell's resultaten het best, in het algemeen waren onze correlaties van dezelfde orde van grootte.

Voor het feit dat de correlaties van de sectie-scores relatief hoog waren in vergelijking met de Kappa's en de discordantiecoëfficiënten zijn twee verklaringen aan te voeren: In de eerste plaats beïnvloeden systematische verschillen tussen beoordelaars de correlatiecoëfficiënten minder dan de Kappa's en de discordantiecoëfficiënten. In de tweede plaats bestaat de mogelijkheid dat beoordelaars die gewend zijn om met complexe eenheden van pathologie te werken, zich wél realiseren dat binnen een bepaald symptoomgebied (sectie) 'iets aan de hand is', doch zich niet al te zeer bekommeren over de vraag bij welk item hun waarneming moet worden ondergebracht.

Bij inspectie van het score-materiaal viel ons dikwijls op dat de beoordelaars wél binnen eenzelfde sectie hadden gescoord, doch verschillende items hadden gebruikt. Deze tendens heeft een nadeliger invloed op de Kappa's en de discordantiecoëfficiënten dan op de overeenstemming voor somscores. Bij de vergelijking van de verschillende onderzoeker-combinaties bleek dat A en C het in de interviewer-observator situaties onderling meer eens waren dan met B.

Wanneer B als interviewer optrad waren de verschillen tussen interviewer en observator wat groter dan wanneer A of C als interviewer optraden. Geen van deze verschillen was echter significant. Wel bleken systematische verschillen in score-hoogte tussen de onderzoekers te bestaan welke statistisch significant waren. Beoordelaar A scoorde vaker positief dan de beoordelaars B en C, terwijl de sectie scores van beoordelaar B lager waren dan die van A en C. De lage drempel van beoordelaar A bij het geven van positieve scores en de neiging van beoordelaar B om in het algemeen lager te scoren dan A en C hingen mogelijk samen met de aard van hun dagelijkse werkzaamheden. Beoordelaar A was werkzaam op een polikliniek waar in het algemeen patiënten met minder ernstige symptomen worden behandeld, terwijl beoordelaar B werkzaam was op een paviljoen waar ernstig gestoorde patiënten werden behandeld en verpleegd. De ernst van de verschijnselen waarmee een beoordelaar in zijn dagelijks werk heeft te maken kan van invloed zijn op de ankerpunten die hij kiest bij het kwantificeren van pathologie.

De veronderstelling vooraf dat de overeenstemming van de interviewer-observator combinaties hoger zou zijn dan van de interviewer-herinterviewer combinaties werd bewaarheid.

Conclusie

Uit deze replicatiestudie is gebleken dat de PSE binnen een klinische setting een geschikt instrument is voor de evaluatie van de psychiatrische status-praesens. Vrijwel alle interviews konden binnen één zitting worden afgenomen. De patiënten leken de procedure niet als hinderlijk te ervaren, en de beoordelaars waren van mening dat de schaal hen een gedegen inzicht in het psychiatrische toestandsbeeld verschaftte.

Uit de relatief ongunstige overeenstemmingsresultaten kunnen een aantal conclusies worden getrokken:

- (1) De bruikbaarheid van de PSE laat te wensen over bij het beoordelen van zeer ernstig gestoorde, met name verwarde of mutistische patiënten. Bij het beoordelen van 80% van de items is men afhankelijk van antwoorden op concrete vragen. De patiënt moet in staat zijn om de bedoeling van de onderzoeker te begrijpen, en er een eenduidig antwoord op geven.
- (2) Een gedeelte van de discordantie hing samen met systematische verschillen tussen beoordelaars, die het gevolg waren van verschillend gekozen ankerpunten. Het verdient aanbeveling om hieraan speciale aandacht te schenken bij de trainingsperiode.
- (3) Een gedeelte van de meningsverschillen tussen de beoordelaars hangt ongetwijfeld samen met het feit dat de beoordelaars onvoldoende getraind waren. Uit de inmiddels elders opgedane ervaring met de schaal is gebleken dat voor het bereiken van een maximale onderlinge overeenstemming een training vereist is, waarbij tenminste 20 patiënten onder supervisie worden beoordeeld. Centra die voornemens zijn om de PSE in te voeren zullen er goed aan

doen om tenminste één lid van hun beoordelaarsteam een door de Medical Research Council georganiseerde trainingssessie te laten bijwonen.

Literatuur

- Ash, P. (1949), The reliability of psychiatric diagnosis, *J. abnorm. soc. psychol.* 44, 272.
- Beck, A. T. (1962), Reliability of psychiatric diagnoses, *Am. J. Psychiatry* 119, 210-216.
- Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educ. psychol. measurement* 20.
- Cooper, J. E., R. E. Kendell, B. J. Gurland a.o. (1972), *Psychiatric diagnosis in New York and London*, Oxford University Press, London.
- 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.
- Kendell, R. E. (1975), *The role of diagnosis in psychiatry*, Blackwell, Oxford.
- Kreitman, N. (1961), The reliability of psychiatric assessment: an analysis, *J. ment. sci.* 107, 887.
- Lin, T. Y. en N. Sartorius (1973), *Report of the international pilot study of schizophrenia*, volume 1, World Health Organization, Geneva.
- Lorr, M. (1966), *Explorations in typing psychotics*, Pergamon, Oxford.
- Lyerly, S. B. (1973), *Handbook of psychiatric rating scales*, 2nd ed., Nat. Inst. of Mental Health, Bethesda, Md.
- Overall, J. E., en C. J. Klett (1972), *Applied multivariate analysis*, McGraw-Hill, New York.
- Rooymans, H. G. M. (1969), *Oordeel en vooroordeel in psychiatrisch diagnostiek*, Proefschrift R.U., Groningen.
- Spitzer, R., J. Cohen, J. L. Fleiss en J. Endicott (1967), Quantification of agreement in psychiatric diagnosis, *Arch. Gen. Psychiatry* 17, 83-87.
- Wing, J. K., J. L. T. Birley, J. E. Cooper, P. Graham en A. D. Isaacs (1967), Reliability of a procedure for measuring and classifying 'present psychiatric state', *Br. J. Psychiatry* 113, 499-515.
- Wing, J. K., J. E. Cooper en N. Sartorius (1974), *The measurement and classification of psychiatric symptoms*, Cambridge University Press.