

Geriatric Mental State Schedule, GMS

Betrouwbaarheidsonderzoek van de Nederlandse editie

*door C. Hooijer, G.C. van der Veer, W. de Rijke en
W. van Tilburg*

[The government] are very keen on amassing statistics – they collect them, add them, raise them to the n -th power, take the cube root and prepare wonderful diagrams. But what you must never forget is that everyone of these figures comes from the ...village watchman, who puts down what he damn pleases. Josiah Stamp, 1929 [citaat uit Blazer 1982].

Samenvatting

De GMS, een gestandaardiseerd interview voor bejaarden, werd in het Nederlands vertaald. De vertaling zelf, de toepassing in Nederland en lacunes in eerder onderzoek vormden de motieven voor het huidige onderzoek. Een betrouwbaarheidsonderzoek werd uitgevoerd. In dit onderzoek bleek de GMS een goed toepasbaar en relevant instrument. De Nederlandse vertaling van de GMS vertoonde dezelfde betrouwbaarheidskenmerken als de Engelstalige versie. De verschillen in de resultaten van eerdere onderzoeken werden grotendeels opgelost. Een eerste onderzoek naar de betrouwbaarheid van de GMS-factoren werd uitgevoerd. Negen factoren bleken voldoende betrouwbaar. De generaliseerbaarheid van deze resultaten werd getoetst en besproken.

Inleiding

De Geriatric Mental State Schedule (GMS) is een semi-gestructureerd, gestandaardiseerd klinisch psychiatrisch interview voor gebruik bij bejaarden. Het instrument werd ontworpen voor het uitvoeren van internationaal vergelijkende studies (Copeland e.a. 1976). De kern bestaat uit 120 standaardvragen. Op basis van de respons van de patiënt op deze vragen en de observatie gedurende het uur dat dit vraaggesprek vergt, worden 548 items gescoord. Uit een deel van deze items kunnen 21 factoren berekend worden (Gurland e.a. 1976), zoals bij voorbeeld 'impaired memory', 'disorientation' of 'anxiety' (voor een volledige opsomming zie ook tabel 7). Deze GMS werd door ons in het Nederlands vertaald en toegepast. De plaats van de GMS in de diagnostiek en ten opzichte van andere instrumenten in de psychiatrie werd elders toegelicht (Hooijer en Van Tilburg 1988). Van de Engelstalige GMS zijn een Brits en een Australisch betrouwbaarheidsonderzoek en een Amerikaans en twee Britse validiteitsonderzoeken bekend (zie par. 2). In dit artikel worden de resultaten van een eigen betrouwbaarheidsonderzoek van de Nederlandse editie besproken.

Motieven voor het Nederlandse betrouwbaarheidsonderzoek

Gestandaardiseerde interviews zijn ontworpen om de verzameling

van gegevens ten behoeve van de psychiatrische diagnostiek meer betrouwbaar te maken. Een algemeen motief om betrouwbaarheidsonderzoek te doen is vast te stellen wat de maximale mogelijkheden van validiteitsonderzoek zijn: als de symptomen van deze patiënt vandaag vastgesteld door deze psychiater niet kunnen voorspellen wat de symptomen van deze patiënt morgen en/of bij een andere psychiater zullen zijn, dan zullen deze symptomen ook niet veel andere zaken kunnen voorspellen. Als betrouwbaarheidsonderzoek ten dienste staat van validiteitsonderzoek dan dient men zich te realiseren dat betrouwbaarheidsresultaten evenzeer gebonden zijn aan de populatie van onderzoek als de validiteitsgegevens. De toepassing van een instrument in een nieuwe (Nederlandse) populatie vormt een *eerste* motief voor herhaald betrouwbaarheidsonderzoek. De vertaling zelf kan zeer heterogene effecten op de uitkomsten van een interview uitoefenen (Spagnoli 1987). Na een dergelijke ingreep moet ten minste aangebond worden dat er nog steeds sprake is van een even betrouwbare methode van dataverzameling.

De voorgaande onderzoeken vertoonden onderling een aantal verschillen, die in de volgende paragraaf toegelicht worden. Deze verschillen vormen op zich een derde motief voor een nieuw (zie derde) betrouwbaarheidsonderzoek van de GMS. Het model dat hiervoor gekozen is is een replicatie van het voorgaand GMS-betrouwbaarheids-onderzoek.

Eerder onderzoek van de Engelse versie

Eerder onderzoek van de betrouwbaarheid van de GMS is uitgevoerd door twee auteursgroepen, de US-UK-groep (Copeland e.a. 1976) en de Australische groep (Henderson e.a. 1983).

a. Copeland e.a. (1976) pasten de GMS toe bij 18 willekeurig gekozen psychiatrisch opgenomen bejaarden in Londen. Zij berekenden een somscore voor het totaal aantal positief bevonden items per patiënt: de produktmomentcorrelaties voor die somscore bedroegen .87 voor een interbeoordelaarsvergelijking en .78 voor een test-hertestvergelijking. Bij vergelijking van de verschillende scores van individuele items bleek een percentage overeenstemming van 93% voor alle vergelijkingen gezamenlijk. De niet-gewogen kappa's per patiënt bedroegen .73 (interrater) en .48 (testretest). Ten behoeve van de kappa-berekeningen per item werden de items geselecteerd op basis van hun frequentie van voorkomen, dat wil zeggen dat in ten minste 13 van de 18 gevallen een van de beoordelaars een item als aanwezig had gescoord. Items waar naar klachten werd gevraagd vertoonden een gewogen kappa-score (Cohen 1968) van .80 en .51 gemiddeld en observatie-items .36 en .29 gemiddeld voor respectievelijk interbeoordelaars- en test-hertestvergelijking.

b. Henderson e.a. (1983) voerden eerst een simulatiestudie uit, ten-

einde de voorwaarden te bepalen waaraan de data moeten voldoen, voordat de berekening van betrouwbaarheidscoëfficiënten per item zinvol mag heten. Alleen de resulterende voorwaarden werden gepubliceerd: (1e) het aantal cases dient minimaal 20 te zijn met (2e) een minimale 'expected cellvalue' van groter dan of gelijk aan 2 en (3e) bij de rekening van de phi-coëfficiënt dient dan het 90% confidentie-interval van de phi kleiner te zijn dan .20 (Bishop e.a. 1976). Het onderzoek vond plaats bij 52 'psychogeriatric outpatients', met behulp van op videocband opgenomen interviews. De phi-coëfficiënten waren dan gemiddeld .83 en .75 respectievelijk voor de interbeoordelaars- en de test-hertestvergelijking.

c. Wanneer we deze twee onderzoeken vergelijken, dan is te zien dat de interbeoordelaarsbetrouwbaarheid, berustend op de vergelijking van de scores tussen een interviewer en een observator, in beide onderzoeken voldoende is. De test-hertestbetrouwbaarheid was bij Copeland e.a. (te) laag maar bij Henderson e.a. goed te noemen. Tevens is te zien dat deze onderzoeken grote verschillen vertonen in onderzoekspopulatie, aantallen en toegepaste itemselectiemethoden en in betrouwbaarheidscoëfficiënten, zodat verschillen in uitkomsten niet eenvoudig te verklaren zijn. De replicatie zal dus zoveel mogelijk beide onderzoeken moeten volgen.

d. Het betrouwbaarheids- en validiteitsonderzoek van de GMS vertonen nog belangrijke lacunes: Het betrouwbaarheidsonderzoek tot nu toe heeft betrekking op een select deel van de 548 individuele items terwijl het validiteitsonderzoek betrekking heeft op de genoemde 21 factoren. Deze factoren worden berekend uit een grotendeels andere selectie van items. Gegevens over de betrouwbaarheid van deze factoren zijn derhalve onbekend. Het lijkt dus zinvol om deze te bepalen. Als vergelijkingsmateriaal kan dan alleen voorgaand validiteitsonderzoek dienen:

e. Eerder validiteitsonderzoek: Fleiss e.a. (1976) passen de 21 factoren toe op een tweedeling van oudere psychiatrische patiënten (intramuraal in New York en Londen) in een subgroep 'organic' en een subgroep 'functional' patiënten. Voor 11 van de 21 factoren werden significante verschillen tussen deze twee subgroepen gevonden. Gebruikmakend van een groep seniel demente patiënten en een groep bejaarde depressieve patiënten vonden Jacoby en Levy (1980) eveneens significante verschillen voor 10 van de 21 factoren. Wanneer we deze twee onderzoeken nu met elkaar vergelijken dan zien we dat voor 9 van de 21 factoren significante verschillen i.c. discriminerende validiteit gevonden is in *beide* onderzoeken, maar dat er voor 12 factoren nog vragen openblijven.

Ons onderzoek van de Nederlandse editie

Voorbereiding – De GMS werd in het Nederlands vertaald door de vier

beoordelaars. Daarbij werd aan de wens van een letterlijke vertaling de voorrang gegeven boven de wens tot vloeiend Nederlands. Drie van hen volgden de GMS-training (Univ. Liverpool, Post-Graduate Course), de vierde werd getraind door de eerste drie.

Methoden – Alle proefpersonen ondergingen tweemaal een interview met een tussentijd variërend van 1 tot 6 dagen. Bij het eerste interview was een observator aanwezig, die geen interventies deed. Deze opzet leidt tot drie GMS-scoringen, te weten van de eerste interviewer (I), de observator (O) en de herinterviewer (H). Vier beoordelaars, destijds psychiaters in opleiding, verrichtten onafhankelijk van elkaar de scoringen van de 55 ppn. Zij fungeerden als I, O en H, wisselend volgens een randomschema, waarbij geen patiënt tweemaal door dezelfde beoordelaar werd gezien. Het onderzoek vond plaats in de middaguren, gedurende de eerste twee weken na opname van de patiënt. De interbeoordelaarsbetrouwbaarheid is gemeten door vergelijking van de scores I en O, de test-hertestbetrouwbaarheid door vergelijking van I en H.

Proefpersonen – De steekproef voor betrouwbaarheid werd samengesteld uit de eerste 55 patiënten en vrijwilligers die geselecteerd werden voor een groter onderzoek betreffende de validiteit van diagnostische methoden voor SDAT (= seniele dementie van het Alzheimer type). De steekproef bestaat daardoor uit klinisch opgenomen patiënten met de diagnose dementie, SDAT (groep D), in dezelfde kliniek opgenomen bejaarde patiënten met psychiatrische gedragsstoornissen waar geen verdenking op enige vorm van dementie bestond (controlegroep C1) en normale vrijwilligers boven de leeftijd van 65 jaar (controlegroep C2). Allen waren behoudens de opnamereden gezond bevonden op basis van uitgebreid lichamelijk onderzoek.

Resultaten

Kenmerken van de steekproef – De verdelingen van diagnose, geslacht, beoordelaars en gemiddelde leeftijd per groep, en beoordelaar zijn gegeven in tabel 1.

Kenmerken van de GMS-scoringen, toepasbaarheid en relevantie – Van de 90.420 mogelijke scores ($55 \times 3 \times 548$) werd 69% ingevuld, de 31% niet ingevulde items betreft 16% legitiem ontbrekende data (niet van toepassing zijnde en facultatieve vragen, e.d.), zodat 15% van het potentiële databestand niet verkregen kon worden. Dit wordt vrijwel geheel veroorzaakt doordat de interviewers op klinische gronden besloten dat informatieverzameling na de eerste GMS-ronde niet zinvol zou zijn. Vrijwel altijd betrof dit patiënten uit de dementiële groep met taalstoornissen, waar in tweede instantie gedurende het interview werd afgezien van taalkundig veeleisende secties als fobie, dwang, depersonalisatie, concentratieverlies e.d.

Van de wel verkregen data bestond 95% van de scores uit de score 0

Tabel 1: Steekproefkenmerken (n = 55)

a. Geslacht, leeftijd en diagnosegroepen

Diagnosegroep:	Aantal	M	V	Leeftijd
D	22	9	13	76,3
C2	13	6	7	74,2
C1	20	8	12	75,1

b. Diagnosegroepen en beoordelaars

Beoordelaars:	A	B	C	D	
Diagnosegroep:					
D	19	15	16	16	66
C2	6	12	10	11	39
C1	19	15	17	9	60
Totalen	44	42	43	36	165

of 1, hetgeen betekent dat dichotomisering geen sterke vergroving inhoudt. Vijfentachtig procent van de scores bedroeg 0 (geen pathologie), 13% wel pathologie in diverse graden (1 t/m 6) en 2% twijfel (8); de verdeling is een uitvloeisel van het brede spectrum van pathologie dat gescoord kan worden. De vraag of niet te veel voor deze groepen irrelevante informatie verzameld wordt lijkt gerechtvaardigd; echter: items die altijd 0 scoorden besloegen 13% van alle items (tegen items die altijd pathologisch scoorden 3%). Gezien de kleine steekproef (n = 55) komt het getal van 13% ons laag voor: ruim 80% van de items komt bij ten minste één patiënt als pathologiescore voor en niet bij allen.

Replicatieonderzoek (de betrouwbaarheid van itemscores)

1. Percentage overeenstemming.

Teneinde de overeenstemming enigszins gewogen te kunnen berekenen zijn de ruwe scores omgezet in een nieuwe schaal naar eigen ontwerp; de schaaldelen zijn gekozen volgens 3 principes:

- de grootste afstand is die tussen géén pathologie en wel pathologie (de scores 0 en 1), deze afstand wordt als eenheid van gebrek aan overeenstemming gekozen;
- deze afstand is gelijk aan het verschil tussen minimum en maximum van de scores betreffende pathologie (1 t/m 6);
- alle andere verschillen zijn hiertussen evenredig verdeeld.

Een overeenstemming van 90% houdt dan in dat van de 55 patiënten er bij hoogstens 5 een dergelijk gebrek aan overeenstemming bestond. Deze overeenstemming van minstens 90% (5 fouten of minder) gold voor 330 (60%) van de items, 4 fouten of minder voor 284 (van de 330!) items, 3 fouten of minder voor 247 items, 2 fouten of minder voor 189 items en één fout voor 135 items. Deze resultaten zijn vergelijkbaar met de 93% overeenstemming gevonden door Copeland.

2. Betrouwbaarheidscoëfficiënten per item

- Van alle items werden de scores omgezet in een dichotomie (0,8 = 0, 1 of meer = 1), gezien het bovenstaande geen grote wijziging.
- Kappa-coëfficiënten werden berekend over het geheel van verschillen en overeenkomsten in verschillende groeperingen: per itemsoort, per diagnosegroep of combinaties daarvan (overall K, tabel 2).

Tabel 2: Kappa berekening over GMS-items in verschillende groeperingen

Betrouwbaarheid: Groep: Itemsoort	Interbeoordelaar				Test-hertest			
	C2	C1	D	Totaal	C2	C1	D	Totaal
Testend	.80	.84	.81	.84	.37	.49	.37	.50
Vragend	.65	.84	.75	.81	.54	.43	.21	.43
Observerend	.62	.62	.64	.65	.31	.35	.38	.40
Geheel	.67	.78	.74	.76	.51	.42	.43	.44

Deze resultaten zijn met betrekking tot de methode min of meer vergelijkbaar met de kappa per patiënt, zoals werd berekend door de groep van Copeland (1976) en de uitkomsten zijn vrijwel identiek, een interbeoordelaarbetrouwbaarheid van .76 en een test-hertestbetrouwbaarheid van .44. Beide eerder vermelde onderzoeksgroepen pasten echter ook verschillende selectiecriteria toe op statistische gronden (zie par. 2). Bij toepassing van beide eerder vermelde sets van criteria werden in ons onderzoek 71 items behouden, waarvan gesteld kan worden dat zij voldoen aan bekende statistische criteria, zodanig dat betrouwbaarheidscoëfficiënten zinvol berekend kunnen worden (tabel 3). De selectiemethoden van beide auteurs heb-

Tabel 3: Replicatie statistische criteria eerdere auteurs in huidig onderzoek

Criteria	niet voldaan	Criteria Henderson	
		niet voldaan	voldaan
Copeland	niet voldaan	–	20
	voldaan	10	71

C= Copeland, 1976: In at least 13 out of 18 cases a symptom is present according to one of the investigators.

H= Henderson, 1983: n of cases ≥ 20 /

minimal expected cell value ≥ 2 / 90% confidence interval of $\phi \leq .2$.

ben blijkbaar een vergelijkbaar effect. Van deze 71 items werden de betrouwbaarheidscoëfficiënten berekend. Voor het overzicht worden hier slechts de gemiddelden weergegeven (tabel 4). De interbeoordelaar betrouwbaarheid was gemiddeld .81 en de test-hertestbetrouwbaarheid gemiddeld .56 (ϕ) en .66 (kappa).

Tabel 4: Betrouwbaarheidscoëfficiënten, GMS, n = 55, gemiddelden. Na toepassing van selectiecriteria, n of items = 71

Interbeoordelaar	Test hertest
kappa .81	kappa .56
phi .81	phi .66

Onze resultaten kunnen nu goed vergeleken worden met eerder onderzoek (tabel 5). De interbeoordelaarbetrouwbaarheid van de twee vorige onderzoeken en het huidige onderzoek zijn niet verschillend. De test-hertestbetrouwbaarheid verschilt wel duidelijk.

Tabel 5: Overzicht vergelijkbare berekeningen betrouwbaarheidsonderzoek GMS

Onderzoeker	n-cases	methode	resultaten			
			interbeoordelaar	test-hertest		
			n-items	coëff.	n-items	coëff.
Copeland	18	k/patiënt	548	.73	548	.48
Huidig onderzoek	55	k/groep	548	.76	548	.44
		selectie				
Copeland	18	Copel. wkap	138	.78	?	.45
Huidig onderzoek	55	Co+He kappa	71	.81	50	.56
Huidig onderzoek	55	Co+He phi	71	.81	50	.66
Henderson	52	Hender.phi	40	.83	14	.75

In volgorde van grootte zijn de gevonden coëfficiënten: Copeland (weighted kappa) .45, dit onderzoek (kappa) .56, dit onderzoek (phi) .66, Henderson (phi) .75. Dat wil zeggen dat daar waar eerdere onderzoekers verschillen, onze resultaten een tussenpositie innemen, die overigens sterk beïnvloed wordt door de keuze van de berekeningsmethode. Ondanks verschillen in methode van berekenen, selecties en patiëntaantallen blijft het belangrijkste fenomeen in deze tabel de relatieve overeenstemming tussen de twee eerdere Engelstalige onderzoeken en het huidige Nederlandse onderzoek.

Deze exercitie diende zoals gezegd om controle op de kwaliteit van de vertaling mogelijk te maken. Voor de validiteitstudies wordt gebruik gemaakt van de factorscores.

Factorscores en betrouwbaarheid

Methode: De factorscores werden berekend door summatie van de betreffende items. Voor elk van de 21 door Gurland e.a. (1976) voorgestelde factoren werd eveneens nu de betrouwbaarheid onderzocht. Proefpersonen, proefopzet en beoordelaars zijn gelijk aan het voorgaande beschreven model: 55 proefpersonen, 3 scoringen verdeeld in 2 vergelijkingen: interbeoordelaar- en test-hertestbetrouwbaarheid. Berekening vindt nu plaats door produkt-momentcorrela-

ties (Pearson's r). Ten behoeve van het overzicht worden de cijfers hier besproken volgens een door ons gekozen indeling van de resultaten in groepen, met een door ons gekozen label. Een overzicht is gegeven in tabel 6.

Tabel 6: Resultaten betrouwbaarheidsonderzoek 21 GMS-factoren ($n = 55$)

Label	Betrouwbaarheidscoëfficiënten *				aantal factoren
	Interbeoordelaar		Test-hertest		
Slecht	0 –	.49	0 –	.49	2
Matig/slecht	.50 –	.74	0 –	.49	1
Discrepant	.75 –	1.00	0 –	.49	9
Voldoende	.75 –	1.00	.50 –	.74	5
Goed	.75 –	1.00	.75 –	1.00	4

* Produktmomentcorrelaties. (Pearson's r)

* Produktmomentcorrelaties. (Pearson's r)

Resultaten:

– Twee factoren voldeden slecht, te weten: 'subjective experience of disordered thought' en 'non social speech'. De eerste is een 'kleine' factor, bestaande uit 3 items, de tweede werd zelden aangetroffen in onze steekproef.

– Eén factor voldeed *matig* tot slecht, te weten "drug and alcoholdependence". Deze factor gaf al in de fase van training en uitvoering aanleiding tot scoringsproblemen. Ten aanzien van deze drie factoren zijn dus veronderstellingen mogelijk, die onbetrouwbaarheid kunnen verklaren.

– Negen factoren vertonen een goede interbeoordelaar- en een slechte test hertest betrouwbaarheid, door ons met het label *discrepant* genoemd. Dit betreft de factoren:

visual hallucinations	retarded speech	hypomania
auditory hallucinations	somatic concerns	observed belligerence
abnormal motor movements	obsessions	reported belligerence

Deze factoren kwamen met uitzondering van 'somatic concerns', weinig voor in onze steekproef. Binnen deze opzet is slechts een beperkte verklaring mogelijk: ofwel het betreft symptomen, die daadwerkelijk fluctueren, ofwel de interview procedure is onbetrouwbaar. Onze proefopzet laat een differentiatie van deze mogelijkheden niet toe.

– Vijf factoren waren *voldoende* betrouwbaar. Dit betrof de factoren 'depression' en 'anxiety', die uit zeer veel items bestaan, en verder de factoren 'disorientation', 'paranoid delusion' en 'depersonalisation and derealisation'.

Vier factoren voldeden *goed*. Dit waren 'impaired memory', 'cortical dysfunction', 'lack of insight' en 'incomprehensibility'.

Validiteit en betrouwbaarheid van de factoren – Hoewel niet eer-

der de betrouwbaarheid van factoren gepubliceerd werd, is toch een indirecte vergelijking mogelijk:

Tabel 7: Validiteit eerder onderzoek en betrouwbaarheid huidig onderzoek

Factor	Validiteit			Betrouwbaarheid		
	Fleiss		Jacoby	Total	Inter	Retest
	New York	c.s. London	c.s. London			
1. Depression	+	+	+	+++	++	+
3. Impaired memory	+	+	+	+++	++	++
12. Disorientation	+	+	+	+++	++	+
21. Incomprehensibility	+	+	+	+++	++	++
6. Somatic concerns	+	+	+	+++	++	-
2. Anxiety	-	+	-	++	++	+
11. Cortical dysfunct.	-	+	+	++	++	++
13. Lack of insight	-	+	+	++	++	++
15. Paranoid delus.	+	-	(+)	++	++	+
10. Drug abuse	-	-	+	+	+	-
18. Audit. hallucinat.	+	-	-	+	++	-
20. Non social speech	+	-	-	+	-	-
14. Depersonalisation	-	-	-	-	++	+
4. Retarded speech	-	-	-	-	++	-
5. Hypomania	-	-	-	-	++	-
7. Observ. belliger	-	-	-	-	++	-
8. Report belliger	-	-	-	-	++	-
9. Obsessions	-	-	-	-	++	-
16. Subj. think. dis.	-	-	-	-	-	-
17. Vis. hallucinations	-	-	-	-	++	-
19. Abn. motor movements	-	-	-	-	++	-

Van de negen factoren waarvoor in beide eerdere validiteitsonderzoeken een diagnostische validiteit werd gevonden blijken er acht overeen te komen met denegen factoren waarvan de betrouwbaarheid in de vorige patragraaf als goed of voldoende werd benoemd. De factoren waarvoor in voorgaand validiteitsonderzoek soms wel en soms niet positieve validiteit werd gevonden, bevinden zich in de groep factoren die in het voorgaande benoemd werden als discrepant wat betreft hun betrouwbaarheidseigenschappen. Vonden wij de betrouwbaarheid slecht of matig, dan werd elders eerder ook geen evidentie voor validiteit gevonden. Uitzonderingen werden gevonden in de factoren 'depersonalisation' (wel betrouwbaar/niet valide) en 'somatic concerns' (niet betrouwbaar/wel valide). De sterke overeenkomst tus-

sen betrouwbaarheidsresultaten en de validiteitsresultaten elders maakt het aannemelijk dat de Nederlandse uitkomsten goeddeels generaliseerbaar zijn over andere intramurale populaties van gedragsgestoorde bejaarden.

Tabel 8: GMS-factoren met voldoende betrouwbaarheid (Pearson's r)

	Interbeoordelaar	Test-hertest
Depression	.97	.73
Anxiety	.88	.57
Paranoid delusions	.98	.69
Depersonalisation/der.	.97	.59
Impaired memory	.95	.89
Cortical dysfunction	.98	.92
Disorientation	.95	.74
Lack of insight	.87	.81
Incomprehensibility	.88	.83

Negen factoren resteren als voldoende betrouwbaar, de gevonden productmomentcorrelaties (Pearson's r) per factor zijn weergegeven in tabel 8. Interrelaties tussen deze negen factoren geven een eenduidig beeld: De laatste vijf, organisch te noemen, factoren vertoonden een sterke samenhang met productmomentcorrelaties boven .70. De eerste vier, functioneel te noemen, factoren vertoonden geen significante samenhang (p .05). Tussen organische en functionele factoren was evenmin enige samenhang.

Voor de overige 12 factoren geldt dat zij in ons onderzoek onvoldoende betrouwbaar bleken te zijn. Voor elf van de actoren (uitzondering 'somatic concerns') is of niet eerder validiteit gevonden of hooguit in één onderzoek wel en in een ander niet. Gezien onze bevindingen hoeven van deze 11 factoren in deze onderzoekspopulaties ook geen gunstige validiteitsuitkomsten verwacht te worden. De veronderstelling dat symptomen sterk fluctueren als verklaring voor de discrepante betrouwbaarheid van 8 factoren, is natuurlijk niet nieuw, maar komt in dit onderzoek onverwacht sterk naar voren. De 8 betreffende factoren doen denken aan een configuratie van bij voorbeeld delirante symptomen. Hierbij dient men te bedenken dat alle proefpersonen zorgvuldig waren onderzocht teneinde somatische aandoeningen van enig belang uit te sluiten.

Discussie

Bij replicatie van eerder onderzoek bleek de Nederlandse editie van de GMS even betrouwbaar als de Engelse versie(s).

Het doel van de berekeningen op itemniveau was controle van de vertaling. Onze kappa-waarden per groep zijn min of meer vergelijk-

baar met Copelands 'kappa per patiënt' (Copeland e.a. 1976). Voor de berekening van betrouwbaarheidscoëfficiënten per item dienen statistische eisen aan de data gesteld te worden. Beide voorgaande onderzoeksgroepen stellen hiertoe criteria voor. Beide sets criteria leken nuttig. Gezien onze replicatie (tabel 3) zijn zij ook redelijk gelijkwaardig. De overeenkomst van onze resultaten met die van eerdere studies is frappant (tabel 5). De betrouwbaarheidscijfers zijn gelijk aan die, welke voor de PSE als voldoende genoemd worden (Duncan Jones e.a. 1980).

De conclusie dat de Nederlandse versie van de GMS even betrouwbaar is gebleken als de Engelstalige lijkt voor de hand te liggen, maar vraagt enig voorbehoud: dit geldt slechts voor dat deel dat controleerbaar was, de 71 items. Een 'breedspectrum' schaal vraagt voor een volledige toetsing ofwel een enorm aantal patiënten met alle mogelijke pathologie, dan wel (per onderdeel) gericht gekozen patiëntgroepen (Henderson 1984).

In onze steekproef was er een verband tussen de aard van deze 71 items en het grote aantal patiënten met de diagnose dementie en de diagnose depressie, de meeste items waren afkomstig uit secties betreffende cognitieve en depressieve symptomen. De selectie op basis van itembetrouwbaarheid zou dus gebruikt kunnen worden als basis voor de constructie van een verkorte schaal van de GMS met betrekking tot dementie en depressie. De geselecteerde items blijken slechts voor een deel te overlappen met de items welke op deze wijze door Henderson e.a. gevonden werden. Dit is een gevolg van enerzijds een verschillende patiëntenpopulatie, i.c. 'psychogeriatric outpatients', waardoor mogelijk andere symptomen in voldoende frequentie voorkwamen en anderzijds doordat Henderson e.a. een versie van de GMS gebruiken, die enigszins afwijkt van de oorspronkelijke US-UK-versie.

Onze resultaten maken bovendien aannemelijk dat het verschil tussen de eerdere onderzoeksresultaten voor een groot deel veroorzaakt wordt door het verschil in gebruikte coëfficiënten.

De betrouwbaarheid van de 21 GMS-factoren werd voor de eerste maal onderzocht. De resultaten laten de volgende grote lijnen zien: factoren betreffende 'organische' fenomenen zijn gewoonlijk betrouwbaar, factoren betreffende 'functionele' fenomenen zijn dat slechts als zij uit veel items bestaan (depression (51) en anxiety (16)). Twee kleinere factoren: paranoid delusions (8) en depers/derealisatie (7) bleken eveneens betrouwbaar zonder dat hiervoor een specifieke verklaring beschikbaar is. Voor de onbetrouwbaarheid van drie factoren is een aannemelijke verklaring mogelijk.

Het meest opvallende gegeven is dat negen factoren in hun scoring sterk wisselden in de tijd van een week. De proefopzet laat geen andere conclusie toe, dan dat het gebruik van deze negen factoren discutabel is en dat zij in validiteitsonderzoek niet zonder meer op eenvoudige

wijze bruikbaar zullen zijn. Gezien de grote gelijkenis van de resultaten van ons betrouwbaarheidsonderzoek van factoren en de resultaten van validiteit van factoren in eerder onderzoek elders, lijkt het aanneemelijk dat bevindingen van het huidige betrouwbaarheidsonderzoek generaliseerbaar zijn voor Engelstalige versies en populaties. Voor validiteitsonderzoek resteren nu negen factoren. Overigens is deze reductie niet zo sterk als zij lijkt: de resterende negen factoren bevatten 116 van de 184 items waaruit de 21 factoren samengesteld waren. Men dient te bedenken dat deze 21 factoren slechts het resultaat zijn van de analyse van één populatie ($n = 100$) (Gurland e.a. 1976), weliswaar met een kruisvalidering op een tweede groep maar daarop nog niet bijgesteld. Het is dan ook niet verwonderlijk dat slechts een deel van de factoren stand houdt in onze populatie.

Gestandaardiseerde interviews zijn ontworpen om de verzameling van data ten behoeve van de psychiatrische diagnostiek meer betrouwbaar te maken. Gewoonlijk betreft dit dan interbeoordelaar- en test-hertestbetrouwbaarheid (Sanson Fisher en Martin 1981). Het begrip betrouwbaarheid kent veel meer aspecten en kent vele valkuilen; begrippen als reactivity of reliability assessment, observerdrift, accuracy versus agreement, proximity of rating (Kazdin 1977) behoren niet tot de gebruikelijke gegevens in de psychiatrie. In feite toetst het gebruikelijke betrouwbaarheidsonderzoek of er inderdaad sprake is van een consensus onder de gebruikers van het instrument. Wij hebben ons hier ook daartoe beperkt. Naast de standaardisering wordt de gebruikelijke manier om consensus te bereiken gevormd door de training van nieuwe gebruikers door (uiteindelijk) de auteur van het instrument.

Wing (1983) noemt zo 20 interviews onder supervisie voor de PSE, een leerproces dat weer onderzocht is door von Cranach en Cooper (1972). De training van de GMS vergt, evenals bij voorbeeld de DIS, 1 week, full-time (Dingemans e.a. 1985). Ook classificatiesystemen, zoals de DSM-III vergen training ook voor psychiaters (Maguire 1982). Over wie getraind moeten/kunnen worden lopen de meningen uiteen: alleen klinici, dat wil zeggen raters met ervaring in het oordelen over psychopathologie of ook leken (Hesselbrock e.a. 1982). Het eerste standpunt kent de meeste aanhang. Training kost tijd (en geld).

In het algemeen vormen cijfers van betrouwbaarheidsstudies een overschatting van de werkelijke betrouwbaarheid van deze data in de dagelijkse praktijk. Deze studies worden uitgevoerd door een selectie van klinici die betrouwbaarheid belangrijk vinden (von Cranach en Cooper 1972). Zij worden vervolgens daartoe getraind, zodat de cijfers omtrent betrouwbaarheid een verdere overschatting vormen van betrouwbaarheid in de dagelijkse praktijk (Grove e.a. 1981).

Een tweede vorm van overschatting is de idee van de universele toepasbaarheid. Schalen zijn getoetst in een bepaalde populatie, die beperkt kan zijn wat betreft diagnose en cultuur. Specifieke schalen zijn

ontwikkeld bij de betreffende diagnosegroepen. Toepassing bij andere diagnosegroepen leidt tot eigenschappen die niet altijd eenvoudig te voorzien zijn (Sanson Fisher en Martin 1981).

Betrouwbaarheidsstudies en hun resultaten dienen dan ook altijd tegen deze achtergrond gezien te worden. Wanneer echter een differentiaaldiagnose, zoals bij voorbeeld het onderscheid dementie versus depressie in de dagelijkse praktijk als moeilijk ervaren wordt, dan kan dit soort onderzoek dienen als een spiegel met de vraag of de in de dagelijkse praktijk toegepaste methoden wel toereikend zijn voor het gevraagde antwoord of dat zij wellicht vervangen moeten worden door deze, zo vaak als 'te kunstmatig' verworpen methode.

Conclusie

De Nederlandse editie van de Geriatric Mental State Schedule kan opgevat worden als een goed toepasbare relevante en betrouwbare methode van dataverzameling ten behoeve van de status presens bij bejaarden met dementiële en niet-dementiële gedragsstoornissen.

Literatuur

- Bishop, Y.M.M., S.E. Fienberg en P.W. Holland (1978), *Discrete Multivariate Analysis: Theory and Practice*, 5e ed., The MIT Press, Cambridge (Mass.).
- Blazer, D. (1982), The Epidemiology of Late Life Depression. *Journal of the American Geriatric Society*, 30, 587-593.
- Cohen, J. (1960), A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20, 37-46.
- Copeland, J.R.M., M.J. Kelleher, J.M. Kellett en A.J. Gourlay met B.J. Gurland, J.L. Fleiss en L. Sharpe (1976a), A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: the Geriatric Mental State Schedule. I. Development and reliability. *Psychological Medicine*, 6, 439-449.
- Cranach, M. von, en J.E. Cooper (1972), Changes in rating behaviour during the learning of a standardized psychiatric interview. *Psychological Medicine*, 2, 373-380.
- Dingemans, P., C. Hooijer en J. Dutilh (1984), ICD-9-CM coderen in de psychiatrie. *Tijdschrift voor Psychiatrie*, 6, 432-444.
- Dingemans, P., H. van Engeland, J.H. Dijkhuis en J. Bleeker (1985), De 'diagnostic interview schedule' (DIS). *Tijdschrift voor Psychiatrie*, 27, 341-259.
- Duncan-Jones, P., S. Henderson en D.G. Byrne (1980), The Present State Examination: Its Use Without Formal Training in a Population Survey. *Social Psychiatry*, 15, 91-93.
- Fleiss, I.L., B.J. Gurland, R. Simon en L. Sharpe (1973), Cross-national study of diagnosis of the mental disorders. Some demographic correlates of Hospital Diagnosis in New York and London. *International Journal of Social Psychiatry*, 19, 180-186.
- Grove, W.M., N.C. Andreasen, P. McDonald-Scott, M.B. Keller en R.W. Shapiro (1981), Reliability Studies of Psychiatric Diagnosis: Theory and Practice. *Archives of General Psychiatry*, 38, 408-413.

- Gurland, B.J., J.L. Fleiss, K. Goldberg en L. Sharpe met J.R.M. Copeland, M.J. Kelleher en J.M. Kellett (1976), A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: the Geriatric Mental State Schedule, II. A factor analysis. *Psychological Medicine*, 6, 451-459.
- Helzer, J.E., L.G.N. Robins, M. Taibleson, R.A. Woodruff, T. Reich en E.D. Wish (1977), Reliability of Psychiatric Diagnosis, A Methodological Review. *Archives of General Psychiatry*, 34, 129-133.
- Henderson, A.S., P. Duncan-Jones en R.A. Finlay-Jones (1983), The reliability of the Geriatric Mental State Examination. *Acta Psychiatrica Scandinavica*, 67, 382-389.
- Hesselbrock, V., J. Stabenau, M. Hesselbrock, P. Mirkin en R. Meyer (1982), A comparison of two interview schedules. The schedule for affective disorders and schizophrenia – lifetime and the National Institute for Mental Health Diagnostic Interview Schedule. *Archives of General Psychiatry*, 39, 674-677.
- Hooijer, C. (1984), De Present State Examination bij bejaarden: de Geriatric Mental State Schedule (GMS). *COBO bulletin*, 17, 9-11.
- Hooijer, C., en W. van Tilburg (1988), Geriatric Mental State Schedule, GMS, een psychiatrisch instrument in de psychogeriatric. *Tijdschrift voor Gerontologie en Geriatrie*, 1988 (subm).
- Jacoby, R.J., en R. Levy (1980), Computed Tomography in the Elderly. 2. Senile Dementia: Diagnosis and Functional Impairment. *British Journal of Psychiatry*, 136, 256-269.
- Kazdin, A.E. (1977), Artifact, Bias and Complexity of Assessment: The ABC's of Reliability. *Journal of Applied Behavioural Analysis*, 10, 141-150.
- Maguire, P. (1982), Psychiatrists also need interview training. *British Journal of Psychiatry*, 141, 423-424.
- Sanson-Fisher, R.W., en C.J. Martin (1981), Standardized Interviews in Psychiatry: Issues of Reliability. *British Journal of Psychiatry*, 139, 138-143.
- Wing, J.K. (1983), Use and Misuse of the PSE. *British Journal of Psychiatry*, 143, 111-117.

Schrijvers zijn respectievelijk: psychiater, UD, vakgroep Psychiatrie, fac. Geneeskunde, VU; psycholoog, vakgroep Psychonomie i.o., sectie Cognitieve Psychologie, subfac. Psychologie en Pedaagogische Wetenschappen, VU; fysi-cus, vakgroep Neurologie, afdeling Klinische Neurofysiologie, fac. Genees-kunde, VU en hoogleraar klinische psychiatrie, vakgroep Psychiatrie, fac. Ge-neeskunde, VU

Corr.: C. Hooijer, Valeriuskliniek, Valeriusplein 9, 1075 BG Amsterdam.

Wij zijn zeer erkentelijk voor de hulp en inzet van onze collega's: Hans van Lieshout, Lex Janssen en Marijke Timmerman.

Het artikel werd geaccepteerd voor publicatie op 17-5-'88.