De OQ is een korte vragenlijst voor het meten van psychische klachten, het interpersoonlijk functioneren en het functioneren in de maatschappelijke rol. Het instrument is bedoeld om bij individuele patiënten het verloop van de klachten over de tijd in kaart te brengen om zo therapie-effect te documenteren. Uit onderzoek komen goede psychometrische kenmerken naar voren. De OQ is ook geschikt voor screening, want zij maakt een adequaat onderscheid tussen patiënten en respondenten uit de bevolking.

Wetenschap

De Outcome Questionnaire OQ-45

Psychodiagnostisch gereedschap

Edwin de Beurs, Margien den Hollander-Gijsman, Victor Buwalda, Wim Trijsburg en Frans Zitman

In de geestelijke gezondheidszorg is het meten van psychopathologie van belang. De overheid en zorgverzekeraars vragen zich af of ze waar krijgen voor het geld dat besteed wordt aan de behandeling van veelvoorkomende psychische problemen, zoals depressie, angst en verslaving. Eén van de manieren om de bestede gelden te verantwoorden, is aantonen dat het beter gaat met een patiënt na een behandeling. Om effect van therapie zichtbaar te maken moet de ernst van de klachten of de kwaliteit van het functioneren van de patiënt worden vastgesteld voor en na de behandeling. Hiertoe worden bij voorkeur gevalideerde meetinstrumenten gebruikt, zoals de Symptom Checklist, (SCL-90, Derogatis, 1975; Arrindell & Ettema, 1986) of de Beck Depression Inventory (BDI, Beck, Ward, Mendelson, Mock & Erbaugh, 1961). Deze instrumenten zijn gericht op het meten van klachten en symptomen. Een gunstige behandelingsuitkomst houdt echter meer in dan alleen symptoomverlichting en kan zich uitstrekken tot gebieden als het functioneren van de patiënt in de partnerrelatie, in het gezin of op het werk. Goed functioneren op deze levensgebieden bepaalt in sterke mate de kwaliteit van leven, maar dit aspect komt in een instrument als de SCL-90 niet aan bod. Alternatieven in de vorm van maten voor de kwaliteit van leven zoals de EuroQoL (EQ-5d, Euroqol Group, 1990) of de RAND36 (of SF-36, Van der Zee & Sanderman, 1993) hebben als bezwaar de grote nadruk op somatische klachten en lichamelijk functioneren.

Recent is door Lambert en collega's (Lambert et al., 1996a; Mueller, Lambert & Burlingame, 1998; Umphress, Lambert, Smart, Barlow & Clouse, 1997) een instrument ontwikkeld dat de ruimte tussen de SCL-90 en de EuroQol kan opvullen: de Outcome Questionnaire (OQ, Lambert, Huefner & Nace, 1997). De OQ is samengesteld als een handzaam instrument met een brede meetpretentie. Het instrument meet naast de emotionele toestand van de patiënt (de aan- of afwezigheid van klachten en symptomen) ook het interpersoonlijke functioneren en het functioneren op maatschappelijk gebied (op het werk of in een opleiding). Er zijn drie subschalen met respectievelijk 25 items voor symptomen, 11 items over problemen met interpersoonlijke relaties en 9 items die disfunctioneren in het werk of in een opleiding representeren (het maatschappelijke rolfunctioneren). Dit brengt het totaal op 45 items, wat de vragenlijst aangenaam kort maakt. Invullen kan in minder dan tien minuten. Het instrument is geschikt bevonden om bijvoorbeeld wekelijks af te nemen zodat veranderingen in het functioneren van de patiënt minutieus gevolgd kunnen worden. Een ander sterk punt van de OQ is dat het instrument zowel items bevat die de *aanwezigheid* van klachten of problemen beschrijven, als ook items die juist een *goede* geestelijke gezondheid of *goed* functioneren in interpersoonlijke relaties of in werk of opleiding beschrijven. Deze laatste groep items dient vanzelfsprekend omgescoord te worden bij de berekening van de totaalscore op de lijst. Het opnemen van positief geformuleerde items ('ik ben gelukkig') is om twee redenen wenselijk: het werkt antwoordtendenties tegen en het is wenselijk omdat het meetinstrument niet alleen de ernst van de klachten moet kunnen vaststellen, maar ook de mate van gezondheid (bijvoorbeeld na succesvolle therapie).

De psychometrische kenmerken van de Amerikaanse versie van de OQ zijn goed (Lambert et al., 1996a; Mueller et al., 1998; Umphress et al., 1997). Chapman (2003) onderzocht tot nog toe de grootste groep respondenten: een steekproef van N = 13.502. Hij rapporteerde Cronbachs α's voor de drie subschalen SD, IR en SR en de totaalscore van respectievelijk $\alpha = 0{,}93$, $\alpha = 0{,}80$, $\alpha = 0{,}74$ en $\alpha = 0{,}94$. De hertestbetrouwbaarheid over een periode van een week was $r = 0{,}80$, $r = 0{,}76$, $r = 0{,}73$ en $r = 0{,}79$. De convergente validiteit werd onderzocht door vergelijking van OQ-schalen met scores op onder meer de SCL-90, de Beck Depression Inventory, de State-Trait Anxiety Inventory en de SF-36 (Umphress et al., 1997). De correlatie van de symptomenschaal van de OQ met de totaalscore op de SCL-90 was $r = 0{,}61$, met de BDI $r = 0{,}63$, met de STAI-trait $r = 0{,}65$ en met de SF-36 $r = 0{,}80$. Met de

totaalscore op de OQ waren deze correlatiecoëfficiënten respectievelijk r = 0,78, r = 0,80, r = 0,80 en r = 0,81. Umphress (1997) toonde ook aan dat op basis van de OQ-totaalscore goed een onderscheid te maken is tussen psychiatrische patiënten en respondenten uit de bevolking.

De OQ is aan een gestage opmars begonnen in de VS en daarbuiten. Zo wordt een 25-itemversie, die alleen items van de symptoom-distressschaal bevat, gebruikt in het ALERT-programma van PacifiCare Behavioral Health, een Californische 'Managed Care'-maatschappij die ggz-voorzieningen biedt aan een populatie van drie miljoen Californiërs. Het ALERT-programma houdt in dat therapie-uitkomstgegevens door de verzekeraar centraal worden bijgehouden en er is een jaarlijkse verkiezing van de meest succesvolle Mental Health Clinic. De Amerikaanse rechthebbende van de OQ heeft de vertaling van de lijst voor een groot aantal taalgebieden ter hand genomen. Zo is er sinds kort ook een Nederlandse versie van de lijst, te verkrijgen via de American Professional Credentialing Services LLC (zie www.oqfamily.com).

Voor een eerste onderzoek naar de betrouwbaarheid en validiteit van de Nederlandse versie van de OQ hebben wij de lijst door drie groepen respondenten laten invullen: eerstejaarspsychologiestudenten, een representatieve steekproef uit de Nederlandse bevolking en een groep ambulante psychiatrische patiënten met voornamelijk stemmings-, angst- en somatoforme stoornissen. In dit artikel worden de psychometrische eigenschappen van de Nederlandse versie van de OQ geëvalueerd. Aan bod komen de betrouwbaarheid (interne consistentie) en de discriminante, convergente en divergente validiteit van het instrument. Ten behoeve van de praktische toepassing van het instrument worden ten slotte normeringsgegevens gepresenteerd voor psychiatrische patiënten en respondenten uit de algemene bevolking. Voor evaluatiedoeleinden bij individuele patiënten worden ook criteria voor betrouwbare verandering en klinische significante verandering gegeven.

Methode

Steekproeven

De OQ werd afgenomen in drie steekproeven: een groep eerstejaarspsychologiestudenten, een representatieve steekproef uit de bevolking en een groep patiënten die zich aanmeldden voor poliklinische behandeling van psychiatrische klachten.

De steekproef psychologiestudenten bestond uit eerstejaarsstudenten van de jaargangen 2001, 2002 en 2003 van de Universiteit van Amsterdam. Zij vulden de OQ-45 in tezamen met een groot aantal andere vragenlijsten in het kader van de 'testweek', een studieverplichting van deze studenten.

De steekproef uit de bevolking was met zorg samengesteld met het oog op representativiteit voor de Nederlandse bevolking als geheel. De steekproef werd getrokken door middel van het telefoonboek en werd gestratificeerd op woonplaatsgrootte en sekse van de respondent, zodat de steekproef in ieder geval op deze twee aspecten representatief zou zijn.[1]

De derde steekproef bestond uit psychiatrische patiënten in ambulante zorg, merendeels patiënten met stemmings-, angst- en somatoforme stoornissen (SAS-stoornissen), die zich aanmeldden voor behandeling op de sectie ambulante volwassenenzorg van de Rijngeestgroep, locatie RijnVeste (Leiden) en locatie RijnAarde (Alphen a/d Rijn) en de Robert Fleury Stichting (locatie Leidschendam).

De OQ

De OQ bestaat uit 45 vragen van het Likerttype. De respondent wordt verzocht aan te geven hoe vaak beschrijvingen op hem/haar van toepassing waren in de afgelopen week, inclusief de dag van invullen. Voorbeelden van items zijn 'Ik was angstig', 'Ik was gelukkig', 'Ik zag de toekomst somber in'. De respondent kan voor de beantwoording kiezen uit vijf antwoordalternatieven: 0 = nooit, 1 = zelden, 2 = soms, 3 = vaak, 4 = bijna altijd.

Naast de totaalscore op de 45 items (TOT) omvat het instrument drie subschalen: ernst van de symptomen of 'Symptom Distress' (SD), interpersoonlijk functioneren of 'Interpersonal Relations' (IR) en maatschappelijk functioneren of 'Social Role' (SR). Bij de SD-schaal voeren depressie en angstklachten de boventoon, maar er zijn ook items opgenomen over middelenmisbruik en -afhankelijkheid. Bij de IR-schaal gaat het om functioneren in relatie tot de partner, eventuele andere gezinsleden, familieleden en vrienden. De SR-schaal ten slotte meet in hoeverre men in staat is maatschappelijke verplichtingen te vervullen op het werk, thuis met huishoudelijk werk of in een opleiding. Conflicten op het werk, een hoge werkdruk ervaren en inefficiënt functioneren op het werk of in een opleiding, leiden tot een hoge score op deze schaal.

Voor de scoring worden items per subschaal opgeteld. Negen items dienen te worden omgescoord (0 = 4, 1 = 3, et cetera). De richting van de schaalscore is telkens zo gekozen dat een hogere score meer symptomen aanduidt, slechtere interpersoonlijke relaties of slechter maatschappelijk functioneren. De range voor de totaalscore loopt van 0 tot 180, voor de SD-schaal van 0 tot 100, voor de IR-schaal van 0 tot 44 en voor de SR-schaal van 0 tot 36.

De SCL-90 en BSI

De Symptom Checklist bestaat uit 90 items van het Likerttype waarin de respondent gevraagd wordt aan te geven in hoeverre hij/zij last had van een klacht of symptoom dat in het item wordt omschreven op een schaal van 1 = 'helemaal niet' tot 5 = 'heel erg'. De 90 items kunnen worden opgeteld tot een totaalscore met een range van 90 tot 450. Daarnaast worden er in de Nederlandse SCL-90 negen dimensies van symptomen onderscheiden. Deze dimensies staan vermeld in de noot onder Tabel 2. De Brief Symptom Inventory (BSI) is een verkorte versie van de SCL-90 van 53 items die tot stand kwam door de Amerikaanse BSI opnieuw te vertalen. De instructies zijn vergelijkbaar met de SCL-90. De items worden beantwoord op een vijfpuntsschaal met de volgende ankerpunten: 0 = helemaal geen, 1 = een beetje, 2 = nogal, 3 =

tamelijk veel, 4 = heel veel. De scores op de BSI worden net zo berekend als in het Amerikaanse origineel: acht dimensies en een totaalscore. Hoe hoger de score, hoe meer klachten.

De SF-36

De Short Form (SF-36) is voortgekomen uit de Medical Outcome Studies (MOS) en bedoeld om de algemene gezondheidstoestand van een patiënt te meten. De SF-36 wordt vaak ingezet als maat voor kwaliteit van leven. Er worden drie aspecten onderscheiden: lichamelijke gezondheid, geestelijke gezondheid en sociale gezondheid (beperkingen in het sociale leven ten gevolge van fysieke of psychische gezondheidsproblemen), gerepresenteerd in acht subschalen (zie noot Tabel 3). De items hebben wisselende antwoordmogelijkheden (zowel ja/nee-items als Likertschalen). Een hogere score duidt op een betere kwaliteit van leven.

Resultaten

De steekproef van eerstejaarspsychologiestudenten

In 2001, 2002 en 2003 werd de OQ afgenomen bij eerstejaarspsychologiestudenten van de faculteit psychologie van de Universiteit van Amsterdam in het kader van de testweek. Dat leverde in 2001 230 ingevulde vragenlijsten op. In 2002 werden 510 vragenlijsten ingevuld en in 2003 completeerden 503 respondenten de OQ. Er waren drie respondenten die alleen nullen of tweeën omcirkeld hadden. Deze zijn uit het bestand verwijderd, zodat er in totaal 1240 respondenten overbleven. De groep bestond voor het merendeel uit vrouwen (882, 71%) en de gemiddelde leeftijd was 21,4 jaar (sd = 5,2; range 17-58).

De steekproef uit de bevolking

Bij het trekken van de bevolkingssteekproef werden 200 volledig ingevulde vragenlijsten geretourneerd, 55% van alle 363 benaderde personen en 78% van de 255 uitgezette vragenlijsten. De responsratio van 78% is voor een onderzoek via de post zeer aanzienlijk. Deelname van 55% van alle benaderde personen maakt representativiteit van de steekproef waarschijnlijk. Desalniettemin werden descriptieve gegevens van de steekproef vergeleken met gegevens van het Centraal Bureau voor de Statistiek over de Nederlandse bevolking anno 2002, het jaar waarin de bevolkingssteekproef werd getrokken. De steekproef bleek representatief wat betreft woonplaatsgrootte, sekse, religieuze achtergrond en arbeidsstatus (werkend, werkloos, WAO). De steekproef verschilde qua leeftijdsopbouw enigszins van de Nederlandse bevolking: de leeftijdsgroep van 18-25 jaar was ondervertegenwoordigd en de leeftijdsgroep van 65 jaar en ouder was iets oververtegenwoordigd. Dit is vermoedelijk veroorzaakt door de telefonische werving via geregistreerde nummers (veel mobiele nummers, alomtegenwoordig onder jeugdige respondenten, komen niet in het telefoonboek voor en werden dus niet in de steekproef betrokken). De gemiddelde leeftijd van de respondenten was 47,5 jaar (sd = 15,0; range 18-88).

De steekproef van patiënten

Patiënten die zich tussen 2002 en medio 2004 hadden aangemeld voor behandeling bij poliklinieken voor volwassenenzorg van de Rijngeest Groep en de Robert Fleury Stichting, ontvingen bij de uitnodiging voor het eerste kennismakingsgesprek vragenlijsten met het verzoek deze thuis in te vullen en ze mee te nemen bij hun eerste afspraak op de kliniek. Aanvankelijk werden zowel de OQ als de SCL-90 bijgevoegd en thuis ingevuld (n = 282). Later werd alleen nog de SCL-90 meegestuurd en werd de OQ op de kliniek elektronisch ingevuld (n = 382). Er werden geen verschillen in OQ-scores gevonden tussen patiënten die de OQ thuis of op de kliniek hadden ingevuld. De groepen zijn om die reden samengenomen. De aldus verzamelde steekproef bestond uit 664 ambulante psychiatrische patiënten. De sekseverdeling in de steekproef was scheef: 420 (63,3%) patiënten waren van het vrouwelijke geslacht. De gemiddelde leeftijd was 36,5 jaar, sd = 11,8, range 17-65.

Bij 387 patiënten in deze steekproef werd een gestandaardiseerd diagnostisch interview afgenomen: de MINI (Sheehan et al., 1998). Bij 356 (91%) werden één of meer DSM-IV-diagnoses vastgesteld, in 316 gevallen (82 %) ging het om een SAS-stoornis. De meest gestelde diagnose was een angststoornis (n = 198), gevolgd door een stemmingsstoornis (n = 176; 169 depressie in engere zin, 6 dystymie en één bipolare stoornis met een huidige depressieve periode). Ten slotte hadden 76 patiënten een somatoforme stoornis. Samen is dit meer dan de 316 patiënten met een SAS-diagnose, hetgeen verklaard wordt door het tegelijk voorkomen van meerdere diagnoses bij één patiënt.

Betrouwbaarheid

In Tabel 1 zijn gegevens opgenomen over de betrouwbaarheid van de OQ-schalen. Als coëfficiënt van interne consistentie werd Cronbachs α berekend. Tevens is voor iedere schaal de gemiddelde itemtotaalcorrelatie en de range opgenomen als alternatieve maat voor de homogeniteit van de schalen. Cronbachs α neemt toe (en is enigszins geflatteerd) naarmate er meer items in een schaal zijn opgenomen. De gemiddelde itemrestcorrelatie is niet gevoelig voor het aantal items, terwijl de 'range' toeneemt met een groter aantal items. De meeste resultaten van Tabel 1 zijn gebaseerd op de gegevens van de patiëntengroep, de meest relevante steekproef voor het instrument. Alleen de testhertestbetrouwbaarheid werd ontleend aan gegevens van 268 eerstejaarsstudenten die in 2002 tweemaal de OQ invulden met een interval van twee weken tussen de twee metingen.

In verband met de beoordeling van de indexen voor interne consistentie is het van belang te vermelden dat voor individuele diagnostische doeleinden meestal een ondergrens van $\alpha = 0{,}80$ wordt aangehouden. Nunnaly en Bernstein (1994) stellen zelfs een kritische grens van 0,90 voor. De SD-schaal en de totaalscore ontstijgen deze grenswaarde en zijn dus voldoende betrouwbaar. De betrouwbaarheid van de IR-schaal is enigszins onder de maat en de SR-schaal is onvoldoende homogeen. Voor de SR-schaal is nog onderzocht of

de interne consistentie stijgt wanneer bepaalde items worden verwijderd. Verwijdering van item 14 doet de α stijgen naar 0,72, verwijdering van item 32 naar 0,73, en verwijderen we ook nog item 28 dan wordt de $\alpha = 0{,}75$. Verdere optimalisatie van de interne consistentie is niet mogelijk.

De testhertestbetrouwbaarheid is met behulp van Pearson productmomentcorrelatiecoëfficiënten geschat bij een deel van de steekproef UvA-studenten uit 2002. 268 studenten vulden de OQ tweemaal in met een tussentijd van veertien dagen. Vanwege extreme antwoorden (bijvoorbeeld alleen scores 0 of 4, drie studenten) of deelname aan een inhaalzitting in plaats van een reguliere zitting (waardoor de tijdsperiode tussen beide invullingen anders is, 11 studenten), resteren in totaal 254 analyseerbare casus. Omdat sommige studenten niet alle vragen invulden, variëren de aantallen per schaal (tussen 235 en 253). Volgens de normen van de Cotan (Evers, Van Vliet-Mulder & Groot, 2000) zijn de testhertestbetrouwbaarheden die zijn weergegeven in Tabel 1 als 'voldoende' aan te merken.

	n	Cronbachs α	R_{it} (range)	R_{tt}
SD	25	0,94	0,37 (-0,02 - 0,74)	0,78
IR	11	0,82	0,28 (-0,01 - 0,55)	0,68
SR	9	0,68	0,18 (-0,28 - 0,53)	0,68
TOT	45	0,95	0,27 (-0,29 - 0,75)	0,78

Noot: OQ: SD = Symptom Distress, IR = Interpersonal Relations, SR = Social Role, TOT = Totaalscore, R_{it} = itemtestcorrelatie; R_{tt} = testhertestcorrelatie

Tabel 1. Aantal items, interne consistentie en itemrestcorrelatiecoëfficiënten van patiënten en testhertestbetrouwbaarheidscoëfficiënten van studenten (14 dagen interval)

De scores over het tijdsinterval werden ook vergeleken met een t-test voor herhaalde metingen. Het is mogelijk dat twee herhaalde metingen hoog correleren terwijl er tegelijkertijd een aanzienlijk verschil is tussen beide meetmomenten, wanneer de scores bij alle respondenten in ongeveer dezelfde mate verschillen tussen de eerste en de tweede afname. De testhertestbetrouwbaarheid van een meetinstrument moet dus niet alleen blijken uit een hoge testhertestcorrelatie, maar ook uit stabiele scores over de tijd. Bij de OQ-schalen was er een gering, maar significant verschil tussen de eerste en de tweede meting. De score op de SD-schaal daalde gemiddeld over het hertestinterval met 3,9 punten, op de IR-schaal 1,4 punten. op de SR-schaal 1,6 schaalpunten en op de totaalschaal 6,9 punten (alle $p < 0{,}001$).

Convergente en divergente validiteit

Convergente validiteit moet blijken uit een hogere samenhang met parallelschalen, divergente validiteit uit een lagere samenhang met andersoortige (sub)schalen. Wil er sprake zijn van voldoende convergente validiteit dan dient de SD-schaal positief samen te hangen met angst- en depressieschalen van de SCL-90 of de BSI, de SR-schaal positief met de SF-36 en de totaalscore positief met de SCL-90- en de BSI-

		SD	IR	SR	TOT
SCL-90	SOM	0,72	0,35	0,34	0,63
	IN	0,72	0,50	0,46	0,70
	SEN	0,67	0,62	0,44	0,69
	DEP	0,81	0,61	0,46	0,80
	ANG	0,76	0,35	0,33	0,67
	HOS	0,55	0,49	0,37	0,57
	AGO	0,62	0,32	0,34	0,57
	SLA	0,59	0,29	0,30	0,55
	TOT	0,87	0,59	0,51	0,85

Noot: alle $p < 0{,}001$

OQ: SD = Symptom Distress, IR = Interpersonal Relations, SR = Social Role, TOT = Totaalscore, SCL-90: SOM = Somatische klachten; IN = Insufficiëntie van denken en handelen; SEN = Interpersoonlijke sensitiviteit; DEP = depressie; ANG = angst; HOS = hostiliteit; AGO = agorafobie; SLA = slaapklachten

Tabel 2. Pearson productmomentcorrelatiecoëfficiënten van de OQ-schalen met de SCL-90, bij een deel van de patiëntensteekproef (n = 320)

		SD	IR	SR	TOT
BSI	SOM	0,57**	0,27**	0,24**	0,49**
	COG	0,68**	0,51**	0,45**	0,68**
	INT	0,61**	0,57**	0,41**	0,65**
	DEP	0,74**	0,62**	0,42**	0,74**
	ANG	0,68**	0,33**	0,31**	0,60**
	HOS	0,53**	0,50**	0,37**	0,56**
	FOB	0,58**	0,32**	0,27**	0,52**
	PAR	0,54**	0,52**	0,37**	0,57**
	PSY	0,65**	0,61**	0,35**	0,67**
	TOT	0,80**	0,59**	0,45**	0,78**
SF-36:	FYS	-0,30**	-0,09**	-0,10**	-0,23**
	SOC	-0,42**	-0,39**	-0,30**	-0,45**
	LIMF	-0,26**	-0,08	-0,17**	-0,22**
	LIME	-0,37**	-0,23**	-0,34**	-0,37**
	MH	-0,71**	-0,54**	-0,45**	-0,72**
	VIT	-0,60**	-0,43**	-0,43**	-0,61**
	PIJN	-0,35**	-0,14**	-0,11*	-0,30**
	ALG	-0,42**	-0,16**	-0,23**	-0,40**

Noot: * $p < 0{,}05$; ** $p < 0{,}01$

OQ: SD = Symptom Distress, IR = Interpersonal Relations, SR = Social Role, TOT = Totaalscore

BSI: SOM = Somatische klachten; COG = Cognitieve problemen; INT = Interpersoonlijke gevoeligheid; DEP = depressie; ANG = angst; HOS = hostiliteit; FOB = Fobische klachten; PAR = Paranoïde gedachten, PSY = Psychoticisme, TOT = Totaalscore

SF-36: FYS =Fysiek functioneren; SOC =Sociaal functioneren; LIMF = Beperkingen t.g.v. lichamelijke problemen; LIME = Beperkingen t.g.v. emotionele problemen; MH = Geestelijke gezondheid; VIT = Vitaliteit; PIJN; ALG = Algemeen gezondheidsbeleven

Tabel 3. Correlaties met de BSI en de SF-36 bij een deel van de patiëntensteekproef (n = 372)

totaalscores. Voor de IR-schaal werd geen parallel instrument ingezet. Divergente validiteit wordt gevonden bij een lagere samenhang van schalen die niet hetzelfde concept meten. De Pearson correlatiecoëfficiënten van de subgroep patiënten die de SCL-90 hadden ingevuld, staan weergegeven in Tabel 2. De resultaten bij patiënten die de BSI en de SF-36 invulden staan in Tabel 3.

In overeenstemming met de verwachting hangen vooral de OQ-SD-schaal en de angst- en depressieschaal van de SCL-90 positief met elkaar samen. Ook de totaalscore van de SCL-90 (psychoneuroticisme) correleert sterk en positief met de OQ-SD-schaal. Een $r = 0{,}87$ komt in de buurt van de bovenlimiet voor convergente validiteit gezien de betrouwbaarheid van beide schalen. De IR- en de SR-schaal laten lagere correlaties zien met de SCL-90-subschalen en de SCL-90-totaalscore. De OQ-totaalscore correleert weer aanzienlijk met de SCL-90, hetgeen onderstreept dat psychopathologie domineert in de OQ. De correlatiecoëfficiënten tussen de BSI en de OQ laten een beeld zien dat met het bovenstaande overeenkomt, al zijn de correlaties over het geheel genomen een fractie lager.

De samenhang met schalen van de SF-36 is over het algemeen lager. De enige SF-36-schaal die aanzienlijk samenhangt met de OQ, is de schaal voor Mental Health. We treffen hier niet de verwachte samenhang van de SF-36 met de OQ-SR-schaal aan. Ook uit Amerikaans onderzoek (Lambert, 1996b) blijkt dat de SF-36 vooral samenhangt met de OQ-SD-schaal en de totaalscore op de OQ (respectievelijk $r_{SD\text{-}MH} = 0{,}80$ en $r_{TOT\text{-}ALG} = 0{,}81$) en in mindere mate met de OQ-IR-schaal ($r_{IR\text{-}SOC} = 0{,}48$).

Ten slotte is de samenhang van de OQ-schalen met de GAF-score onderzocht. De score op de GAF representeert zowel de ernst van de psychiatrische klachten als het niveau van functioneren. Er werden derhalve geen differentiële voorspellingen gedaan voor de verschillende OQ-schalen. De GAF bleek vooral samen te hangen met de OQ-SD-schaal (correlaties bedroegen respectievelijk $r_{SD\text{-}GAF} = 0{,}35$, $r_{IR\text{-}GAF} = 0{,}20$, $r_{SR\text{-}GAF} = 0{,}22$ en $r_{TOT\text{-}GAF} = 0{,}34$). Deze correlatiecoëfficiënten tussen een beoordelingsschaal en de zelfrapportage zijn lager dan wat gevonden werd bij de vergelijking van subschalen van telkens twee zelfrapportage-instrumenten, wat verklaard wordt door het verschil in methode van dataverwerving.

Discriminante validiteit

Een goede test voor discriminante validiteit is of psychiatrische patiënten en respondenten uit de normale bevolking verschillend scoren op de OQ. Psychologiestudenten werden niet gebruikt als vergelijkingsgroep, omdat deze groep op veel andere variabelen van patiënten verschilt, zoals opleidingsniveau en levensomstandigheden. Overigens is de gemiddelde score van studenten vergelijkbaar met die van normalen (zie Tabel 4 voor schaalgemiddelden van de drie groepen). De scores op de vier schalen van de OQ van patiënten en respondenten uit de bevolkingssteekproef werden onderzocht met een viertal t-tests voor onafhankelijke steekproeven. De uitkomsten van de t-test staan weergegeven in Tabel 4. Er zijn forse verschillen tussen patiënten en respondenten uit de bevolking, oplopend tot bijna twee standaardeenheden voor de symptoomschaal en de totaalscore. Met deze schalen is het derhalve goed mogelijk een onderscheid te maken tussen patiënten en de bevolkingssteekproef. We hebben dit verder onderzocht met discriminantanalyse. De resulterende discriminantfunctie is significant (Wilks $\Lambda = 0{,}62$, $\chi^2(3) = 317{,}17$, $p < 0{,}001$). In de discriminantfunctie domineert de SD-schaal met een Standardized-Discriminant-Function-coëfficiënt van 0,80, gevolgd door de IR-schaal (0,16) en de SR-schaal (0,15). Met informatie van de drie subschalen kan 92% van de patiënten en 70% van de gezonde respondenten als zodanig worden geclassificeerd. Met een stapsgewijze procedure kan nog worden onderzocht wat de zuinigste set van subschalen is om onderscheid te maken tussen beide groepen. De SD en de IR blijken dan te volstaan. Toevoeging van de SR-schaal leidt niet tot een significante verbetering van het onderscheidend vermogen. De twee schalen classificeren 91% van de patiënten en 67% van de gezonden correct. Dezelfde analysetechniek is toegepast om het onderscheidend vermogen van de OQ tussen studenten en patiënten te onderzoeken. Ook dan vinden we een significante discriminantfunctie (Wilks $\Lambda = 0{,}65$, $\chi^2(3) = 749{,}73$, $p < 0{,}001$), waarin de SD-

	Studenten (1240)		Bevolking (200)		Patiënten (664)		Vergl. Patiënten Bevolking	
	x	sd	x	sd	x	sd	T(863)*	Cohens d
SD	25,27	12,98	22,80	12,32	48,54	15,86	23,3	1,81
IR	10,42	5,71	8,80	5,38	16,62	6,64	16,4	1,29
SR	9,86	3,80	7,79	3,31	13,41	5,17	17,0	1,29
TOT	45,60	20,24	39,23	18,73	77,72	23,77	22,7	1,80

*Alle $p < 0{,}001$

Noot: OQ: SD = Symptom Distress, IR = Interpersonal Relations, SR = Social Role, TOT = Totaalscore

Tabel 4. Gemiddelde score van patiënten en respondenten uit de bevolking

	Optimale **sensitiviteit**			Optimale **specificiteit**			
schaal	Cut-off	sens.	spec.	cut-off	sens.	spec.	AUC
SD	32	0,86	0,78	37	0,77	0,87	0,89
IR	12	0,74	0,75	13	0,66	0,84	0,82
SR	10	0,77	0,71	12	0,63	0,86	0,82
TOT	52	0,87	0,77	65	0,70	0,88	0,89

Noot: OQ: SD = Symptom Distress, IR = Interpersonal Relations, SR = Social Role, TOT = Totaalscore

Tabel 5. Grensscores, sensitiviteit en specificiteit voor OQ-schalen

schaal op de voorgrond staat met een coëfficiënt van 1,20. Met deze functie worden 59,6% van de patiënten en 92,0% van de studenten juist geclassificeerd.

Met een Receiver-Operator-Characteristic-analyse (ROC-analyse) kan worden onderzocht wat het optimale afkappunt is voor iedere schaal om de overgang van normale klachten naar psychopathologie te markeren. In Tabel 5 staan voor de OQ schalen de optimale afkappunten, de bijbehorende sensitiviteit en specificiteit en de 'area under the curve' (AUC, een index van het onderscheidend vermogen van de schalen met een range van 0,5 tot 1,0) berekend met een ROC-analyse van de gegevens van de patiëntensteekproef en de respondenten uit de bevolking. Deze gegevens representeren het vermogen van de schalen om een onderscheid te maken tussen de respondenten uit de normale bevolking en de poliklinische patiënten. We zijn er hierbij dus vanuit gegaan dat geen van de respondenten uit de bevolkingssteekproef aan een psychiatrische stoornis lijdt. Deze veronderstelling zal niet geheel juist zijn: de prevalentie van psychiatrische stoornissen wordt geschat op 10 tot 15% van de bevolking (Nemesisonderzoek (Bijl, Van Zessen & Ravelli, 1997). De resultaten zijn dus iets vertekend ten nadele van de OQ.

Verschillen tussen mannen en vrouwen in score op de OQ

Sekseverschillen in score op de OQ zijn onderzocht om na te gaan of er verschillende normen en grensscores voor mannen en vrouwen gehanteerd zouden moeten worden. Lambert et al (1996b) melden dat er met de Engelstalige versie geen sekseverschillen gevonden werden. De vergelijking van scores van mannen en vrouwen wijzen uit dat alleen op de SD-schaal er bij alledrie de steekproeven een significant verschil is. Het verschil is niet groot en bedraagt bij de bevolkingssteekproef 4.6 schaalpunten (= 0.4 standaardafwijking). Bij de studenten vinden we tevens verschillen op de IR-schaal en de OQ-totaalscore, maar deze verschillen zijn klein (< 1/6de standaardafwijking). Het verschil in score is echter wel zo groot dat er verschillende normen voor mannen en vrouwen nodig zijn.

Criteria voor klinisch significante verandering bij individuele patiënten

Door Jacobson en medewerkers (Jacobson, Follette & Revenstorf, 1986) zijn criteria opgesteld voor klinisch significante verandering. Wil er sprake zijn van een klinisch betekenisvolle verandering, dan moet aan twee criteria voldaan worden: (1) statistisch betrouwbare verandering ('Reliable Change' of RC: een zo grote verschuiving in score dat de kans kleiner is dan 5% dat deze verschuiving berust op onnauwkeurigheid van het instrument en (2) overschrijding van een grenswaarde (Cut-Off of CO) die de overgang markeert van ziek naar gezond (herstel) of omgekeerd (terugval). Wordt alleen het eerste criterium gehaald dan is er sprake van betrouwbare verbetering of verslechtering, maar (nog) niet van herstel of terugval. Wordt alleen het tweede criterium gehaald dan is er weliswaar een verschuiving van ziek naar gezond of omgekeerd, maar bevinden beide scores zich zo dicht bij de grensscore dat die verschuiving (nog) geen klinische betekenis heeft. De RC- en CO-waarden voor de OQ staan in Tabel 6. Gezien het sekseverschil dat werd aangetroffen op de SD- en de totaalscore van de OQ zijn er verschillende grensscores voor mannen en vrouwen.

	Nederlandse versie			Amerikaanse versie*	
	RC	CO		RC	CO
		mannen	vrouwen		
SD	11	32	36	10	36
IR	8	13	13	8	15
SR	7	11	10	7	12
TOT	15	56	57	14	63

RC = Reliable Change; CO = Grenswaarde voor overgang van pathologisch naar gezond

NB.: De RC- en CO-waarden in de tabel doen zelf mee: Bij de SD-schaal moet de verschilscore 11 of groter zijn en een score van 35 of meer is pathologisch.

Tabel 6. Indices voor betrouwbare verandering

De grensscores voor klinische verandering zijn bij de Amerikaanse versie iets hoger en de RC-index is iets lager dan bij de Nederlandse versie. Nederlandse patiënten moeten dus iets meer veranderen van voor- naar nameting om te kunnen spreken van klinisch significante verbetering. De verschillen zijn echter gering.[2]

Discussie

Alvorens de bevindingen kritisch tegen het licht te houden, is het goed een paar beperkingen van de huidige studie onder de aandacht te brengen. Ten eerste bestond de onderzochte patiëntengroep voor een aanzienlijk deel uit patiënten die waren voorgeselecteerd op de aanwezigheid van een stemmings- angst- of somatoforme stoornis. Weliswaar is hiermee het grootste deel van de patiënten die zich aanmelden voor ambulante zorg in de ggz meegenomen, maar bepaalde groepen (bijvoorbeeld patiënten met verslavingsproblematiek) zijn niet vertegenwoordigd. De OQ kan een bruikbaar instrument zijn in de verslavingszorg, maar hoe dergelijke patiënten scoren, weten we dus vooralsnog niet. Ook is de bruikbaarheid van de OQ bij opgenomen psychiatrische patiënten niet onderzocht. Voorts is de studie beperkt tot de analyse van baselinegegevens van de patiëntengroep, waardoor de gevoeligheid van de OQ voor therapie-effecten nog niet aan bod is gekomen. Dat de OQ bijzonder goed patiënten van normalen kan onderscheiden, houdt een belofte in voor gevoeligheid voor therapie-effecten die we in een vervolgstudie zullen onderzoeken. Ten slotte schiet het design van de huidige studie tekort bij de bepaling van de convergente validiteit voor de subschaal voor interpersoonlijke relaties van de OQ. De convergente validiteit van de IR-schaal dient dus nog nader onderzocht te worden. Goede kandidaten voor dit doel zijn de Interactioneel Probleem Oplossend Vermogen (IPOV, Lange, 1983) of de Dyadic Adjustment Schaal (DAS, Spanier, 1982). De SF-36 als parallelschaal voor maatschappelijk functioneren is ook niet de optimale keuze. In dit instrument ligt de nadruk sterk op beperkingen in het functioneren vanwege somatische problemen. De Social Adjustment Self-report Scale (SAS, Weissman & Bothwell, 1976) zou beter geschikt zijn om de concurrente validiteit van de SR-schaal te onderzoeken, maar is helaas een lange lijst.

Ondanks deze beperkingen is een goede eerste indruk verkregen van de sterke en de zwakke kanten van de Nederlandse versie van de OQ. De betrouwbaarheid in termen van interne consistentie van de subschalen blijkt ruim voldoende. De testhertestbetrouwbaarheid is iets lager dan gewenst, maar daarbij tekenen we aan dat deze alleen is vastgesteld bij de steekproef van studenten. Bij de vaststelling van de stabiliteit van de OQ-scores bij deze groep over tijd trad iets opmerkelijks op: de gemiddelde score was lager bij de tweede afname, zij het in geringe mate. Verlaging van de score bij herhaalde afname kan een kenmerk zijn van de OQ (als de respondent dit instrument voor een tweede keer ziet is hij/zij geneigd lager te scoren op de items), maar het kan ook een accurate weerspiegeling van verandering in de steekproef zijn, die door de OQ wordt opgepikt, gevoelig als het instrument is voor verandering. Verandering van score bij herhaalde afname is overigens geen unieke eigenschap van de OQ. Ook andere meetinstrumenten voor psychopathologie laten dat vaak zien (vergelijk Arrindell, 1993). Zo werd bij paniekstoornispatiënten een afname van 0,30 standaarddeviatie in de score op de SCL-90 in een wachtlijstperiode van drie maanden aangetroffen (De Beurs, 1993).

Bij de samenstelling van de OQ zijn items geselecteerd om drie onderling samenhangende gebieden van functioneren te meten: last van symptomen (of juist het ontbreken daarvan), tevredenheid of problemen in interpersoonlijke relaties en tevredenheid of problemen in het functioneren in de samenleving, bijvoorbeeld op het werk, opleiding of met vrijetijdsbesteding. De OQ is in de eerste plaats bedoeld als een meetinstrument om verandering ten gevolge van behandeling te meten. De drie gebieden zijn gekozen vanuit de overtuiging dat verandering zich moet uitstrekken over die gebieden en niet beperkt moet blijven tot een afname van de kernsymptomen van een aandoening (Lambert, Hansen & Finch, 2001). De 'face'-validiteit van de items en schalen is goed. De validiteit van de OQ wordt in het algemeen ook ondersteund door de resultaten van dit onderzoek. Dat we bij de SD-schaal te maken hebben met een goede maat voor psychopathologie, mag blijken uit de sterke samenhang met de SCL-90 en haar kortere variant, de BSI. De discriminante validiteit van de OQ wordt ondersteund door het onderscheid dat het instrument kan maken tussen patiënten enerzijds en de normale bevolking en psychologiestudenten anderzijds. Het blijkt een goede screener voor psychopathologie. Getuige de resultaten van de stapgewijze discriminantanalyse is de SD-schaal het best geschikt om een onderscheid te maken tussen patiënten en gezonde respondenten. Op de IR- en de SR-schaal verschillen beide groepen minder. Therapie-effect zal naar verwachting ook vooral op de SD-schaal tot uiting komen.

Conclusie

Een groot voordeel van de OQ-45 is dat het een korte lijst is met een simpele en eenduidige opzet. Ondanks het relatief kleine aantal items bestrijkt de OQ een breder gebied van disfunctioneren dan bestaande klachtenlijsten. Uit de resultaten komen goede psychometrische eigenschappen naar voren: de subschalen zijn voldoende betrouwbaar en de validiteit van de symptomenschaal en de totaalscore wordt ondersteund. Al met al rechtvaardigen de resultaten tot nu toe nader onderzoek naar de OQ dat zich zal richten op twee aspecten: de factoriële structuur van de lijst (vormen de drie voorgestelde subschalen de optimale onderverdeling van de items) en de gevoeligheid voor therapie-effect in vergelijking met andere veelgebruikte meetinstrumenten.

Dr. E. de Beurs is verbonden aan de Afdeling Psychiatrie van het Leids Universitair Medisch Centrum, en aan de Rivierduinen.
E-mail: <E.de_Beurs@lumc.nl>.
Mw drs. M. den Hollander-Gijsman is verbonden aan de Afdeling Psychiatrie van het Leids Universitair Medisch Centrum.
Drs. Victor Buwalda is verbonden aan de Rivierduinen.
Prof.dr. W. Trijsburg is verbonden aan de Programmagroep Klinische Psychologie, Universiteit van Amsterdam en de Afdeling Medische Psychologie & Psychotherapie, ErasmusMC, Roterdam.
Prof.dr. F. Zitman is verbonden aan de Afdeling Psychiatrie van het Leids Universitair Medisch Centrum, en aan de Rivierduinen.

Noten

Een uitgebreide versie van dit artikel met aanvullende normeringsgegevens is beschikbaar op <http://www.lumc.nl/3010/algemeen/OutcomeQuestionnaire.pdf>.

1. De auteurs danken dhr J. van der Plas en dhr J. Dijkgraaf voor het verzamelen van de gegevens in het kader van hun onderzoeksstage.
2. Bij de berekening van de RC is in navolging van Lambert et al. (1996b) gebruik gemaakt van de interne consistentie als betrouwbaarheidsindex van het instrument. Meer gebruikelijk is het evenwel om hiervoor de testhertestbetrouwbaarheid te gebruiken. In het algemeen valt die lager uit dan Cronbachs α (0,70/0,80 vs. 0,80/0,90). Lambert et al. gebruiken echter de Cronbachs α als betrouwbaarheidsindex, waarmee de geschiktheid van het instrument om verandering over de tijd te meten er gunstiger uitziet dan op basis van testhertestbetrouwbaarheid.

Literatuur

Arrindell, W.A. (1993). The fear of fear concept: stability, retest artefact and predictive power. *Behaviour Research and Therapy*, *31*, 139-148.

Arrindell, W.A. & Ettema, J.H.M. (1986). *SCL-90: Handleiding bij een multidimensionele psychopathologie-indicator*. Lisse, The Netherlands: Swets & Zeitlinger.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J.E. & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561-571.

Beurs, E. de (1993). *The assessment and treatment of panic disorder with agoraphobia*. Academisch proefschrift, Universiteit van Amsterdam. Amsterdam: Thesis Publishers.

Bijl, R.V., Zessen, G. van & Ravelli, A. (1997). Psychiatrische morbiditeit onder volwassenen in Nederland: het NEMESIS-onderzoek. II Prevalentie van psychiatrische stoornissen. *Nederlands Tijdschrift voor de Geneeskunde*, *141*, 2453-2460.

Chapman, J.E. (2003). *Reliability and validity of the Progress Questionnaire: an adaptation of the Outcome Questionnaire*. Philadelphia, PA: Drexel University.

Derogatis, L.R. (1975). *The Symptom Checklist-90-R*. Baltimore, MD: Clinical Psychometric Research.

Euroqol Group (1990). Euroqol – a new facility for the measurement of health-related quality of life. *Health Policy*, *16*, 199-208.

Evers, A., Vliet-Mulder, J.C. van & Groot, C.J. (2000). *Documentatie van tests en testresearch in Nederland*. Amsterdam/Assen: NIP/Van Gorcum.

Jacobson, N.S., Follette, W.C. & Revenstorf, D. (1986). Toward a standard definition of clinically significant change. *Behavior Therapy*, *17*, 308-311.

Lambert, M.J., Burlingame, G.M., Umphress, V., Hansen, N.B., Vermeersch, D.A., Clouse, G.C. et al. (1996a). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy*, *3*, 249-258.

Lambert, M.J., Hansen, N.B. & Finch, A.E. (2001). Patient-focused research. Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, *69*, 159-172.

Lambert, M.J., Hansen, N.B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G.M. et al. (1996b). *Administration and scoring manual for the OQ-45.2*. Wilmington, DE: American Professional Credentialing Services LLC.

Lambert, M.J., Huefner, J.C. & Nace, D.K. (1997). The promise and problems of psychotherapy research in a managed care setting. *Psychotherapy Research*, *7*, 321-332.

Lange, A. (1983). *De Interactionele Probleem Oplossingsvragenlijst, IPOV*. Deventer: Van Loghum Slaterus.

Mueller, R.M., Lambert, M.J. & Burlingame, G.M. (1998a). Construct validity of the outcome questionnaire. A confirmatory factor analysis. *Journal of Personality Assessment*, *70*, 248-262.

Nunnally, J.C. & Bernstein, I.R. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E. et al. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59 Suppl 20*, 22-33.

Spanier, G.B. (1982). Measuring dyadic adjustment. New scales for assessing quality of mariage and similar dyads. *Journal of Marriage and the Family*, *38*, 15-28.

Umphress, V.J., Lambert, M.J., Smart, D.W., Barlow, S.H. & Clouse, G. (1997). Concurrent and construct validity of the outcome questionnaire. *Journal of Psychoeducational Assessment*, *15*, 40-55.

Weissman, M.M. & Bothwell, S. (1976). Assessment of social adjustment by patients self-report. *Archives of General Psychiatry*, *33*, 1111-1115.

Zee, K.I. van der & Sanderman, R. (1993). *Het meten van de algemene gezondheidstoestand met de RAND-36, een handleiding*. Rijksuniversiteit Groningen: Noordelijk Centrum voor Gezondheidsvraagstukken.

Summary

The Outcome Questionnaire (OQ-45): measuring psychiatric symptoms and interpersonal functioning

E. de Beurs, M. den Hollander-Gijsman, V. Buwalda, W. Trijsburg, F. Zitman

The OQ is a short self-report questionnaire for the assessment of psychiatric symptoms, interpersonal relations en social role functioning. Thus, the instrument measures a more comprehensive set of variables than a symptom checklist alone, such as the SCL-90. The instrument is intended for patient-focused research, assessing individual progress during treatment. The present research investigates the reliability and validity of the OQ. The OQ was administered in three groups of respondents: first-year psychology students, a representative sample from the general population, and psychiatric outpatients. The results showed the psychometric properties to be good: reliability was sufficient and validity was supported by sufficient concordance with other self-report measures. The instrument is suitable for assessing the symptoms in outpatients. Furthermore, the instrument discriminates well between patients and respondents from the general population, which makes the OQ suitable for screening purposes. For practical purposes, cut-off scores for reliable change and the transition from a pathological to a healthy state are presented.