De Nederlandse versie van de Outcome Questionnaire (OQ-45): een crossculturele validatie

Article *in* Psychologie & gezondheid · February 2008

DOI: 10.1007/BF03077465

2 authors:

Kim de Jong
Leiden University
46 PUBLICATIONS **1,071** CITATIONS
SEE PROFILE

Philip Spinhoven
Leiden University
535 PUBLICATIONS **27,517** CITATIONS
SEE PROFILE

Kim de Jong, Annet Nugter , Marike Polak, Hans Wagenborg[1], Philip Spinhoven en Willem Heiser

De Nederlandse versie van de Outcome Questionnaire (OQ-45): een crossculturele validatie[2]

Kim de Jong en Annet Nugter zijn werkzaam bij de Afdeling onderzoek, GGZ Noord-Holland-Noord, Heiloo. Kim de Jong is tevens werkzaam bij de Universiteit Leiden, evenals Marike Polak, Philip Spinhoven en Willem Heiser. Philip Spinhoven is tevens verbonden aan het Leids Universitair Medisch Centrum.

Correspondentieadres: Mw. drs. K. de Jong, GGZ Noord-Holland-Noord, afdeling Onderzoek, Postbus 18, 1850 BA Heiloo, *E-mailadres*: k.dejong@ggz-nhn.nl

SUMMARY

The Dutch version of the Outcome Questionnaire (OQ-45): a cross-cultural validation

The cross-cultural validity of the OQ in the Dutch population has been examined by comparing the psychometric properties and equivalence in factor structure and normative scores of the Dutch OQ with the original American version. Data were collected at university ($N = 268$), in community ($N = 810$) and in three mental health care organizations ($N = 1920$). Results showed that the psychometric properties of the Dutch OQ were adequate and similar to the original instrument. Some differences in equivalence were found though. In factor analysis, two additional factors were found: one consisting of Social Role items and another that reflected anxiety and somatic symptoms. Furthermore, normative scores were different for the Dutch and American samples and this resulted in different cut-off scores for estimating clinically significant change in the Dutch population.

Inleiding

De Outcome Questionnaire (OQ; Lambert et al., 1996) is de laatste jaren steeds populairder geworden in onderzoek naar behandelresultaten en is inmiddels vertaald in meer dan vijftien talen. Hoewel de psychometrische eigenschappen van de originele versie van de OQ grondig onderzocht zijn, zijn er weinig publicaties beschikbaar over de kenmerken van vertaalde versies. Dit artikel gaat over de crossculturele validatie van de Nederlandse OQ.

De reden voor de stijgende populariteit van de OQ ligt in het feit dat het instrument enkele eigenschappen heeft die de meeste andere instrumenten

1 Dr. Hans Wagenborg, afdeling onderzoek en ontwikkeling van de Geestgronden, is kort na de acceptatie van het originele artikel overleden.

2 Dit artikel is een bewerking van Jong, K. de, Nugter, M.A., Polak, M.G., Wagenborg, J.E.A., Spinhoven, P. & Heiser, W.J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: a cross-cultural validation. *Clinical Psychology & Psychotherapy, 14*, 288-301.

niet hebben. In de eerste plaats meet de OQ meerdere domeinen van psychosociaal functioneren. Het is tegenwoordig gebruikelijk bij uitkomstonderzoek naast symptoomreductie ook de verbetering in het algemeen functioneren te meten (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999). Andere populaire uitkomsteninstrumenten zoals de SCL-90 (Derogatis, 1977), de *Brief Symptom Inventory* (BSI; Derogatis, 1975) of de *Social Adjustment Scale* (SAS; Weissman & Bothwell, 1976) meten of alleen de symptomen of alleen het functioneren. Verder is de OQ een algemeen instrument dat gebruikt kan worden voor meerdere stoornissen, waardoor het mogelijk is om een vergelijking te maken tussen de behandeluitkomsten van een grote verscheidenheid aan patiënten, onafhankelijk van de psychiatrische diagnose. Specifieke instrumenten zijn ontwikkeld voor specifieke stoornissen en kunnen dus niet de symptomen en het functioneren van verschillende soorten patiënten vergelijken. Bovendien is de OQ relatief kort. De patiënt kan hem in ongeveer 5 minuten invullen.

Het belangrijkste voordeel van de OQ is dat de vooruitgang van patiënten gevolgd kan worden met herhaalde metingen. De OQ wordt vaak gebruikt bij onderzoek waarbij wekelijkse feedback gegeven wordt aan de behandelaar over de vooruitgang van de patiënt. Het verloop van de behandeling van de patiënt wordt vergeleken met een voorspeld verloop en de therapeut wordt gewaarschuwd wanneer de patiënt te veel van het voorspelde verloop afwijkt. Deze feedback leidt tot effectievere behandelingen, vooral voor die patiënten die het risico lopen onvoldoende van de behandeling te profiteren (zie o.a. Lambert, 2007) . De OQ kan dan dus ook relevant zijn voor procesonderzoek van behandelingen.

In het vertalingproces kunnen de eigenschappen van een instrument veranderen. Het kan zijn dat de betekenis van sommige items verandert in een andere taal. Ook kunnen de onderliggende constructen anders zijn in een andere cultuur. Culturele verschillen beïnvloeden de *conceptuele equivalentie* van twee verschillende talenversies van een instrument (Flaherty et al., 1988). Een methode om de conceptuele equivalentie te onderzoeken is het vaststellen of de factorstructuur vergelijkbaar is in de nieuwe taalversie. Een bekend voorbeeld van een gebrek aan conceptuele equivalentie is de SCL-90: de Nederlandse versie heeft een andere factorstructuur dan de originele versie (Arrindell & Ettema, 1975).

Er bestaan vaak verschillen tussen de normatieve scores van verschillende culturen, zelfs tussen westerse landen. Een verschil in normatieve scores kan leiden tot een verschillende criteriumvaliditeit. Dit kan de gevoeligheid en specificiteit van het instrument veranderen. Flaherty et al. (1988) noemen dit een gebrek aan *criteriumequivalentie*. Bij onderzoek naar klinische uitkomsten is een gebruikelijk criterium het feit of er een klinisch significante verandering (Jacobson & Truax, 1991) heeft plaatsgevonden tijdens de behandeling. Omdat de grenswaarde of *cut-off* score voor klinische significante verandering gebaseerd is op de bevolkingscurven, is het belangrijk om goede steekproeven te nemen in de nieuwe cultuur. Als er geen criteriumequivalentie bereikt wordt, moeten de scores gekalibreerd worden en moeten er dus nieuwe *cut-off* scores berekend worden (Flaherty et al, 1988).

Het doel van dit onderzoeksproject was om de equivalentie van de Nederlandse en de Amerikaanse OQ vast te stellen en om de psychometrische kenmerken van de Nederlandse OQ te bepalen. Eerder onderzoek heeft aangetoond dat Nederlanders lagere scores hebben op de OQ dan Amerikanen (De Jong & Nugter, 2004). Ook werden er enkele verschillen gevonden bij de psychometrische kenmerken. Omdat de grootte van de normatieve steekproef nog onvoldoende was voor individueel gebruik van de OQ en er meer informatie nodig was over de validiteit, zijn er meer gegevens verzameld. Voor dit artikel zijn de oude en nieuwe gegevens samengevoegd.

Methode

De data

De data werden verzameld bij een steekproef van studenten psychologie, twee steekproeven uit de normale populatie en twee klinische steekproeven. Een groep van 268 studenten vulden een OQ, SCL-90 en GVSG-45 in en 264 studenten vulden twee weken later de OQ voor de tweede keer in. Deze gegevens werden gebruikt om de betrouwbaarheid en validiteit te berekenen en niet voor normering, omdat deze groep niet voldoende representatief is voor de normale populatie.

Een steekproef van 448 personen uit de normale populatie werd verkregen door willekeurig personen te selecteren uit 13 telefoonboeken, geografisch verdeeld over heel Nederland. Alle volwassenen van een gezin werden gevraagd om de OQ in te vullen. De respons bij deze groep was 55%. Een tweede steekproef van 362 personen werd verkregen via 14 bedrijven in verschillende sectoren. Hier was de respons 33%. Bij de steekproef uit de normale populatie zijn 24 personen die onder behandeling waren voor psychologische of psychiatrische problemen buiten beschouwing gelaten.

De eerste klinische steekproef van 1545 ambulante patiënten werd verzameld binnen drie GGZ-instellingen in Noord-Holland. Binnen de steekproef zaten zowel patiënten uit stedelijk als uit landelijk gebied. Alle patiënten die zich aanmeldden in de onderzoeksperiode vulden de OQ in voor of na het intakegesprek. Om de test-hertestbetrouwbaarheid vast te stellen, vulde een subgroep van 43 patiënten de OQ twee tot drie weken later opnieuw in; tussentijds vond geen behandeling plaats. Een andere subgroep van 117 patiënten vulde naast de OQ ook de SCL-90 en DASS in. Met behulp van een internet screeningstool (Interapy Nederland B.V., Amsterdam) werd nog een steekproef getrokken van 375 patiënten die de OQ online invulden voor de intake.

Een overzicht van de kenmerken van de steekproeven wordt gegeven in Tabel 1. In de normale populatie zijn enkele verschillen tussen de substeekproeven gevonden, omdat één van de bedrijven, een thuiszorgorganisatie, relatief veel vrouwen onder haar werknemers had. De steekproef van dit bedrijf werd toch meegenomen omdat de scores voor de OQ niet significant anders waren dan bij de andere bedrijven. Om technische redenen waren de demografische kenmerken van de internet- selectiesteekproef niet beschikbaar.

Tabel 1. Kenmerken van de data.

Steekproef	*N*	Sekse		Leeftijd	
		Vrouw *N (%)*	Man *N (%)*	*Bereik*	*(SD)*
Normale populatie	810	513 (63)	297 (37)	18-94	44.3 (15)
– Steefproef uit telefoonboek	448	248 (55)	200 (45)	18-94	47.9 (16)
– Bedrijven steekproef	362	265 (73)	97 (27)	18-77	39.4 (12)
Studenten	268	171 (64)	96 (36)	17-53	22.3 (6)
Klinische steekproef					
– Ambulante patiënten, papieren afname	1545	896(58)	628(41)	18-65	37.3 (11)
• test-hertest	42	31 (74)	11 (26)	18-55	31.7 (10)
• concurrent validiteit SCL-90/DASS	118	76 (64)	41 (35)	18-61	33.6 (11)
• gevoeligheid voor verandering	60	32 (53)	24 (40)	23-62	41.6 (10)
– Ambulante patiënten, internet screening tool	375	-	-	-	-

Er werden 8 studenten, 4 personen uit de normale populatie en 49 patiënten verwijderd uit de steekproef, omdat ze meer dan 20% van de vragen op een vragenlijst niet beantwoordden. Voor de overige ontbrekende waardes werden de schaalscores berekend op basis van de resterende antwoorden op de schaal, mits ten minste 80% van de antwoorden op de schaal waren beantwoord. Ontbrekende waardes werden niet vervangen in analyses op itemniveau, zoals factoranalyse en betrouwbaarheidsanalyse.

Instrumenten

De Outcome Questionnaire. De OQ bestaat uit 45 items waarop gescoord kan worden op een schaal van vijf punten, van *nooit* (0) tot *bijna altijd* (4). De Symptomatische Distress (SD) subschaal bestaat uit 25 items die betrekking hebben op de meest voorkomende psychiatrische stoornissen, zoals depressie, angst en drank- en drugsverslaving. De Interpersoonlijke Relaties (IR) subschaal bestaat uit 11 items en meet het functioneren van de patiënten in relaties met hun partner, familie en vrienden. De Sociale Rol (SR) subschaal bevat 9 items en meet het functioneren op school, werk en in de vrije tijd.

Instrumenten gebruikt voor validatie van de OQ. In deze sectie wordt een korte beschrijving gegeven van de instrumenten die gebruikt zijn om de OQ te valideren. Alle instrumenten zijn zelfinvullijsten en hebben voldoende goede psychometrische eigenschappen[1] .

De *Symptom Checklist* 90-item versie (SCL-90; Arrindell & Ettema, 1975) en *Depression Anxiety and Stress Scales* (DASS; De Beurs, Van Dyck, Marquenie, Lange, & Blonk, 2001) zijn gebruikt om de SD subschaal te valideren. Bij de SCL-90 is de *Global Severity Index* (GSI) berekend, de gemiddelde score over alle items. Bij de DASS zijn de schaalscores van de Depres-

sie, Angst en Stress subschalen gecorreleerd met de OQ schaalscores.

De Groningse Vragenlijst Sociaal Gedrag 45-item versie (GVSG-45; De Jong & Van der Lubbe, 2001) is gebruikt om de IR en SR subschalen te valideren. De GVSG-45 meet sociaal gedrag op 9 gebieden. Voor dit onderzoek zijn in overleg met de auteurs van de GVSG-45 twee nieuwe indices berekend die niet in de oorspronkelijke vragenlijst zitten. Als index voor interpersoonlijke problemen is het gemiddelde over de Ouders, Partner, Kinderen en Vrienden subschalen berekend; verder Functioneren op Interpersoonlijke Relaties (FIR) genoemd. Als index voor de sociale rol is een gemiddelde over de schalen School, Werk, Huishoudelijk werk en Vrije tijd berekend; hierna Functioneren op de Sociale Rol (FSR) genoemd.

Data analyse

Voor alle toetsen voor verschillen tussen groepen werd een significantieniveau van $\alpha = 0.05$ gehanteerd. Effectgroottes (Cohen's *d*) werden berekend met de spreadsheet van Thalheimer en Cook (2002) in Microsoft Excel en werden geïnterpreteerd volgens de criteria van Cohen (1992).

De indices waarover gerapporteerd wordt in de confirmatieve factoranalyse zijn de *Root Mean square Residual* (RMR), de *Root Mean Square Error of Approximation* (RMSEA), de *Goodness-of-Fit Index* (GFI), de *Bentler-Bonnet Normed Fit Index* (NFI), *de Comparative Fit Index* (CFI), de chi-square (χ^2) en de chi-kwadraat gedeeld door het aantal vrijheidsgraden (χ^2/df). Algemene richtlijnen geven aan dat de RMR kleiner dan 0.10 zou moeten zijn, de RMSEA kleiner dan 0.05, de GFI groter dan 0.95, de CFI and NFI groter dan 0.90, een χ^2 die niet significant is en χ^2/df kleiner dan 2 (Kline, 1998).

Resultaten

Conceptuele equivalentie

De steekproef bestaande uit de gecombineerde data van de normale en klinische steekproeven werd aselect in tweeën gesplitst, zodat wanneer de analyses een afwijkende factorstructuur zouden tonen voor de Nederlandse OQ de stabiliteit van de nieuwe structuur getoetst kon worden op de tweede helft van de gegevens. De eerste helft van de data werd gebruikt om de drie-factorstructuur van de OQ te toetsen. Er werd een confirmatieve factoranalyse met de *generalised least squares* (GLS) schattingsmethode uitgevoerd.

Tabel 2. Confirmatieve factoranalyse Goodness of Fit Indices.

Model	χ^2	*df*	χ^2/df	RMR	RMSEA	GFI	NFI	CFI
Drie-factoroplossing	3678.2	942	3.90	.103	.046*	.880	.933*	.949*
Vijf-factoroplossing	3413.4	925	3.69	.075*	.044*	.889	.957*	.964*

RMR = Root Mean-square Residual; RMSEA = Root Mean-Square Error of Approximation; GFI = Goodness-of-Fit Index; NFI = Normed Fit Index; CFI = Comparative Fit Index.

* Voldoet aan het aanbevolen criterium

Zoals te zien is in Tabel 2, voldoet onze drie-factoroplossing aan drie van de zeven *goodness-of-fit* criteria. De indices die niet voldeden aan de *goodness-of-fit* criteria waren de χ^2 , χ^2/df, RMR en GFI. De χ^2 en χ^2/df zijn afhankelijk van de grootte van de steekproef en, gezien de grote omvang van onze steekproef en de daaruit voortvloeiende grote *power*, is een significante χ^2 niet noodzakelijkerwijs een aanwijzing van slechte *goodness-of-fit*. De RMSEA, NFI en CFI voldoen aan de criteria voor een goede *goodness-of-fit* en de drie-factoroplossing lijkt een redelijke *fit* te hebben. De *goodness-of-fit* van onze oplossing is duidelijk beter dan de oplossing die Mueller, Lambert en Burlingame (1998) verkregen hebben; hun drie-factoroplossing voldeed aan geen van de criteria die wij toegepast hebben.

Tabel 3 toont de gestandaardiseerde factorladingen voor de drie factoren. Vier items hebben factorladingen die lager zijn dan 0.30: item 11, 14, 26 en 32. Items 11, 26 en 32 staan bekend als moeilijke items. Ze meten alle drie problematische drank- en/of drugsgebruik en hebben een scheve scoreverdeling (veel '0' scores). Item 14 'Ik werk/studeer te veel' is een ander speciaal geval. Dit item scoort ook niet goed in de originele OQ en heeft een negatieve correlatie met meerdere items in de covariantiematrix. Bovendien is het het enige item in

Tabel 3. Gestandaardiseerde factorladingen voor de factormodellen.

	3-factoroplossing (*N* = 1362)			5-factoroplossing (*N* = 1363)				
Item	F1	F2	F3	F1	F2	F3	F4	F5
2	.66			.65				.16
3	.63			.62				
5	.71			.73				
6	.73			.72				
8	.67			.66				
9	.85			.82				.13
10	.78			.67				.38
11	.18			.14				
13	.84			.87				
15	.84			.87				
22	.68			.68				
23	.80			.83				
24	.75			.77				
25	.76			.75				.14
27	.45			.36				.34
29	.61			.51				.44
31	.86			.88				
33	.71			.61				.32
34	.37			.33				.39
35	.48			.38				.39
36	.73			.68				.29
40	.73			.71				
41	.57			.56				.25
42	.84			.87				
45	.45			.41				.34
1		.59			.60			
7		.60			.60			
16		.45			.41			.21
17		.46			.44			
18		.81			.81			
19		.58			.62			
20		.71			.76			
26		.27			.27			
30		.74			.73			
37		.70			.71			
43		.80			.84			
4			.71			.51	.39	
12			.63			.68		
14			.13			-.11	.34	
21			.76			.73		
28			.34			.30		
32			.21			.20		
38			.81			.66	.50	
39			.65			.44	.48	
44			.66			.58		

de SR subschaal waar de normale steekproef ($M = 1,87$, $SD = 1,1$) zelfs iets hoger scoort dan de klinische steekproef ($M = 1,68$, $SD = 1,2$), $t(1706) = 3.54$, $p < 0.001$, $d = 0.14$.

Omdat de drie-factoroplossing nog niet bevredigend was, hebben we geprobeerd om meer variantie te verklaren door de residuele matrix van de drie-factoroplossing te gebruiken. Een principale componentanalyse met varimax rotatie leverde twee componenten op met een eigenwaarde hoger dan 1, die samen 34%van de variantie verklaarden. Voor elke component werden de items gekozen met een lading groter dan 0.15.

De twee componenten werden toegevoegd aan het originele drie-factormodel en toegepast op de andere helft van de steekproef. De χ^2/df had een verbeteringsratio van 15.6, wat als een aanzienlijke verbetering beschouwd kan worden. De andere waarden voor de *goodness-of-fit* zijn eveneens verbeterd. Vooral de RMR-waarde is belangrijk, die voldeed niet aan het criterium van 0.10 in de drie-factoroplossing, maar wel in de vijf-factoroplossing.

De eerste extra factor bestaat uit items die allemaal behoren tot de Sociale Rol subschaal. De niet-verklaarde variantie in de SR subschaal lijkt voornamelijk veroorzaakt te zijn door item 14, die een lage factorlading heeft op dit domein. Voor klinische doeleinden heeft de extra sociale factor weinig toegevoegde waarde. Het kan van pas komen als iemand alleen geïnteresseerd is in het functioneren in de sociale rol, maar in dat geval zou de OQ waarschijnlijk niet het instrument van voorkeur zijn. Daarom is ervoor gekozen deze factor verder niet mee te nemen in de analyses.

De tweede extra factor lijkt wel iets toe te voegen aan de klinische bruikbaarheid van het instrument. De items zijn voornamelijk afkomstig uit de SD subschaal en lijken verband te houden met cognitieve (bijvoorbeeld 'Ik ben angstig') en somatische (bijvoorbeeld 'Mijn hart klopt te snel') representaties van angst. Daarom hebben we besloten om deze nieuwe factor wel mee te nemen in de verdere analyses. Deze factor wordt verder Angst en Somatische Distress (ASD) genoemd.

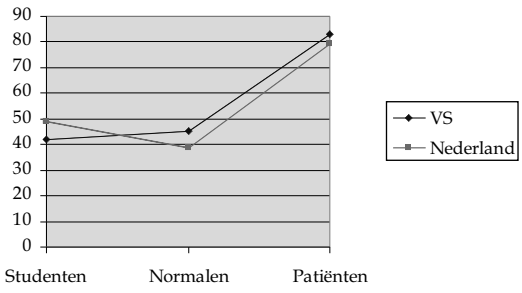
Criteriumequivalentie

Verschillen in scores. Tabel 4 geeft de gemiddelde scores van de Nederlandse en Amerikaanse steekproeven weer op de OQ subschalen en de totale schaal. De gemiddelde totaalscores voor de Nederlandse steekproeven onder de normale ($t(1622) = 7.48$, $p < 0.001$, $d = 0.37$) en klinische populatie ($t(2261) = 2.50$, $p = 0.01$, $d = 0.15$) liggen iets onder de overeenkomstige Amerikaanse scores, hoewel de effectgroottes klein zijn. De gemiddelde scores van de Nederlandse studenten zijn iets hoger dan die van de Amerikaanse studentensteekproef ($t(502) = -4.40$, $p < 0.001$, $d = 0.39$). Figuur 1 geeft een visuele weergave van de verschillen tussen de steekproeven voor de totale schaal van de OQ.

In de Amerikaanse steekproeven waren er geen verschillen tussen mannen en vrouwen. In de Nederlandse steekproeven zijn enkele verschillen aangetroffen. Bij de normale populatie werden er significante verschillen gevonden tussen de seksen, Wilks' $\lambda = 0.92$, $F(5, 792) = 13.3$, $p < 0.001$. Vrouwen hadden hogere niveaus van Symptomatische Distress, $F(1, 796) = 10.7$, $p = 0.001$, $d = 0.26$, en Angst en Somatische Distress, $F(1, 796) = 21.8$, $p < 0.001$, $d = 0.37$, terwijl mannen meer problemen hadden met het functioneren in de Sociale Rol, $F(1, 796) = 13.6$, $p < 0.001$, $d = 0.27$.

Soortgelijke resultaten werden gevonden in de klinische steekproef: Wilks' $\lambda = 0.92$, $F(5,1446) = 23.7$, $p < 0.001$. Ook hier hadden vrouwen iets hogere niveaus van Symptomatische Distress, ($F(1,1450) = 5.83$, $p = 0.016$, $d = 0.13$) en Angst en Somatische Distress ($F(1,1450) = 29.1$, $p < 0.001$, $d = 0.29$) terwijl mannen wat meer problemen hadden met het functioneren in de Sociale Rol ($F(1,796) = 25.4$, $p < 0.001$, $d = 0.27$).

Figuur 1. OQ-totaalscore voor de Amerikaanse en Nederlandse populaties.



Klinische significantie en betrouwbare verandering

Om de individuele verandering te meten wordt vaak het criterium van klinische significantie gebruikt. Het criterium is tweevoudig: (a) de grootte van de verandering moet statistisch betrouwbaar zijn en (b) aan het eind van de behandeling moeten de patiënten in een scorerange vallen die hen niet onderscheidt van normaal functionerende personen. Een *cut-off* score voor klinisch significante verandering en een *reliable change* index zijn hier uitgerekend met behulp van de c-formule van Jacobson en Truax (1991).

Tabel 4. Gemiddelden en standaarddeviaties van de OQ in Nederlandse en Amerikaanse data.

	Amerikaanse gegevens						Nederlandse gegevens					
	Studenten (*N* = 235)		Normalen (*N* = 815)		Patiënten (*N* = 342)		Studenten (*N* = 268)		Normalen (*N* = 807)		Patiënten (*N* = 1920)	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Symptomatische Distress	23.0	10	25.4	12	49.4	15	27.3	12	22.2	10	48.9[b]	16
Angst en Somatische Distress	-	-	-	-	-	-	15.6	7	13.3	6	25.9[b]	9
Interpersoonlijke Relaties	8.8	5	10.2	6	19.7	6	11.4	5	8.4	5	16.8[c]	7
Sociale Rol	10.4	4	9.6	4	14.1	5	10.4	4	8.1[a]	3	13.6[d]	6
OQ-totaalscore	42.2	17	45.2	19	83.1	22	49.1	18	38.7	16	79.5	25

Opmerking: Gemiddelden en standaardafwijkingen van de Amerikaanse data zijn overgenomen uit de Amerikaanse handleiding van de OQ (Lambert et al, 2004).

a N = 798; b N = 1919; c N = 1917; d N = 1849.

Voor de Nederlandse OQ is de *cut-off* score 55 voor de totale schaal en 12 voor de IR subschaal. Gezien de verschillen die gevonden zijn tussen de mannelijke en vrouwelijke respondenten, werden voor de subschalen met sekseverschillen aparte *cut-off* scores voor mannen en vrouwen berekend. De *cut-off* score voor de SD subschaal was 31 voor mannen en 33 voor vrouwen, voor de ASD subschaal was dit 17 voor mannen en 19 voor vrouwen en voor de SR subschaal was het 12 voor mannen en 10 voor vrouwen. Iemand die een score heeft die hoger of gelijk is aan de *cut-off* score behoort tot de disfunctionele (klinische) range.

Bij een *cut-off* score van 55 is de gevoeligheid voor de totale schaal van de OQ 0.84, wat betekent dat 84% van de steekproef van de normale populatie correct geïdentificeerd is als behorend tot de functionele steekproef. De specificiteit van de OQ is 0.85, wat betekent dat 85% van de klinische steekproef correct geïdentificeerd is als disfunctionerend.

De *reliable change* index (RCI) wordt uitgedrukt als het aantal punten op een bepaald meetinstrument dat de patiënt moet verbeteren tijdens de behandeling om van betrouwbare verbetering te kunnen spreken. De RCI is afhankelijk van de betrouwbaarheid van het meetinstrument en de variabiliteit van de scores. Als betrouwbaarheidsindex werd de interne consistentie van de klinische en normale steekproeven samen gebruikt (zie Tabel 5). De RCI's voor de SD, ASD, IR en SR subschalen zijn respectievelijk 10, 8, 8 en 9. De RCI voor de OQ totale schaal is 14.

Psychometrische kenmerken

Betrouwbaarheid

De interne consistentie van de subschalen en de totale schaal is voldoende tot goed (zie Tabel 5), behalve voor de SR subschaal, waarvoor teleurstellende Cronbach's α's werden gevonden. De meeste waardes komen overeen met de waarden die in de Amerikaanse steekproef gevonden werden. Er zijn geen waardes van de interne consistentie in de Amerikaanse normale populatie bekend, maar in een Duitse steekproef onder de normale populatie was Cronbach's α voor de SR subschaal 0.59, wat dicht bij onze waarde van 0.53 komt (Lambert, Hannöver, Nisslmüller, Richard, & Kordy, 2002).

De test-hertestbetrouwbaarheid is een aanwijzing voor de stabiliteit van de vragenlijst. De correlatie tussen de eerste en de tweede keer dat de OQ ingevuld werd is voldoende tot goed voor zowel de klinische (r_{tt} = 0.70-0.83) als de studentensteekproef (r_{tt} = 0.71-0.81).

Tabel 5. Interne consistentie (Cronbach's α) and test-hertestbetrouwbaarheid (Pearson's product-moment correlation coefficient).

	Interne consistentie								Test-hertest	
	Studenten		Normale populatie		Patiënten		Patiënten en normalen		Studenten (*N* = 264)	Patiënten (N = 42)
Schalen	*N*	α	*N*	α	*N*	α	*N*	α	*r*	*r*
Symptomatische Distress	257	0.90	768	0.89	1247	.91	2390	0.95	0.81	0.76
Angst en Somatische Distress	261	0.79	786	0.82	1743	0.84	2529	0.89	0.74	0.70
Interpersoonlijke Relaties	264	0.74	770	0.77	1607	0.80	2377	0.84	0.71[a]	0.83
Sociale Rol	258	0.61	773	0.53	1620	0.69	2393	0.72	0.73	0.74
OQ-totaal score	247	0.92	726	0.91	1309	0.93	2035	0.96	0.82	0.79

a N = 262

Validiteit

Criteriumvaliditeit. Een belangrijke vereiste voor een meetinstrument dat behandeluitkomsten meet is dat het onderscheid kan maken tussen de klinische populatie waarvoor hij bestemd is en de functionele (normale) populatie. Tabel 4 laat zien dat de verschillen tussen de gemiddelden van de normale populatie en de klinische populatie groot zijn. De steekproef onder de normale populatie functioneert significant beter dan de klinische populatie op alle subschalen en op de totale schaal, Wilks' $\lambda = 0.58$, $F(5, 2637) = 388.1$, $p < 0.001$. Effectgroottes voor de verschillen tussen de klinische steekproef en de steekproef onder de normale populatie zijn zeer groot voor alle subschalen (SD: $F(1, 2643) = 1873.2$, $p < 0.001$, $d = 1.83$; ASD: $F(1, 2643) = 1280.7$, $p < 0.001$, $d = 1.52$; IR: $F(1, 2643) = 960.5$, $p < 0.001$, $d = 1.32$; SR: $F(1, 2643) = 674.1$, $p < 0.001$, $d = 1.10$) en de totale schaal ($F(1, 2643) = 1804.5$, $p < 0.001$, $d = 1.80$).

Concurrente validiteit. De SCL-90 en DASS werden gebruikt om de SD en ASD subschalen te valideren (zie Tabel 6). De concurrente validiteit van de SD subschaal met de GSI van de SCL-90 was iets lager dan de Amerikaanse waarde bij de klinische steekproef ($r = 0.80$ versus $r = 0.84$), maar beter bij de studentensteekproef (0.78 versus 0.61). De correlaties tussen de SD subschaal en de DASS subschalen waren toereikend. De ASD subschaal heeft voldoende concurrente validiteit met de SCL-90 en de Angstsubschaal van de DASS ($r = 0.74$). De correlaties tussen de Depressie ($r = 0.63$) en Stress ($r = 0.60$) subschalen en de ASD-subschaal waren lager, zoals verwacht.

Het was moeilijk om een Nederlands instrument te vinden om de IR en SR subschalen te valideren en we moesten onze eigen indices uitrekenen voor het instrument dat we uiteindelijk gebruikt hebben. Toch valt de convergente validiteit van de GVSG-45 met de IR ($r = 0.51$) en SR subschaal ($r = 0.55$) in de range van correlaties van de OQ subschalen met respectievelijk de *Inventory of Interpersonal Problems* (IIP) (r = 0.49-0.64) en de *Social Adjustment Scale* (SAS) in de Amerikaanse steekproeven (r = 0.44-0.73). Echter, de samenhang van de IR subschaal met de GVSG-45 is niet erg specifiek, de samenhang is bijna net zo hoog met de FIR als met de FSR index.

Gevoeligheid voor verandering

Een ander belangrijk criterium voor uitkomstinstrumenten is dat ze veranderingen in het functioneren moeten kunnen meten die het gevolg zijn van de behandeling. Een subgroep van 60 patiënten onderging een korte behandeling, van maximaal vijf sessies. De OQ werd voor en na de behandeling afgenomen. De OQ toonde een grote gevoeligheid voor veranderingen op alle subschalen (SD: $t(55) = 6.8$, $p < 0.001$, $d = 1.29$; ASD: $t(56) = 7.7$, $p < 0.001$, $d = 1.43$; IR: $t(51) = 4.3$, $p < 0.001$, $d = 0.84$; SR: $t(55) = 4.1$, $p < 0.001$, $d = 0.77$) en de totale schaal, $t(56) = 7.1$, $p < 0.001$, $d = 1.33$.

Tabel 6. Concurrent validiteit voor de OQ met de SCL-90, DASS en GVSG-45.

	Patiënten (*N* = 118)				Studenten (*N* = 268)		
	GSI	DASS-D	DASS-A	DASS-S [a]	GSI	FIR	FSR
Symptomatische Distress	**0.80**	**0.78**	**0.74**	**0.72**	**0.78**	0.42	0.54
Angst en Somatische Distress	**0.75**	0.63	**0.74**	0.60	**0.66**	0.34	0.42
Interpersoonlijke Relaties	0.62	0.54	0.38	0.54	0.59	**0.51**	0.51
Sociale Rol	0.51	0.51	0.46	0.48	0.57	0.38	**0.55**
OQ-totaal score	0.80	0.77	0.68	0.72	0.77	0.49	0.60

GSI = Global Severity Index van de Symptom Checklist 90 – Revised (*scl*-90-R)

DASS-D = Depression Anxiety Stress – Depression subscale

DASS-A = Depression Anxiety Stress – Anxiety subscale

DASS-S = Depression Anxiety Stress – Stress subscale

FIR = Functioneren in Interpersoonlijke Relaties, gebaseerd op de Groningse Vragenlijst Sociaal Gedrag 45-item versie (GVSG-45) subschalen Ouders, Partner, Kinderen en Vrienden

FSR = Functioneren in de Sociale Rol, gebaseerd op de GVSG-45 subschalen School, Werk, Huishoudelijk werk en Vrije tijd

Alle correlaties zijn significant op $p < .01$

a N = 117

Discussie

In dit onderzoek werd de crossculturele validiteit van de OQ in de Nederlandse bevolking onderzocht, door de psychometrische kenmerken van de Nederlandse versie en de equivalentie met de originele versie van de OQ te beoordelen. De resultaten laten zien dat de twee versies overeenkomen in betrouwbaarheid en validiteit, maar er werden verschillen gevonden in de factorstructuur en de normatieve scores.

De drie-factorstructuur van het instrument, waarvoor geen sterk bewijs was in de originele versie, had een redelijke *fit* in de Nederlandse populatie. Nadere analyses leverden twee aanvullende factoren op. De eerste extra factor, die bestond uit vier items die tot het domein van de Sociale Rol behoren, was onverwacht maar begrijpelijk gezien het slechte functioneren van item 14 ('Ik werk/studeer te veel'). Dit item is ook problematisch in de originele OQ en vertegenwoordigt waarschijnlijk geen problematisch gedrag. In de hedendaagse maatschappij zullen sommige mensen te hard werken zelfs als een goede eigenschap beschouwen.

De tweede factor, de Angst en Somatische Distress, werd beschouwd als een nuttige aanvulling op de bestaande schalen en werd daarom meegenomen in de verdere analyses. De betrouwbaarheid- en validiteitschattingen voor de ASD-factor waren veelbelovend. Deze factor kan vooral interessant zijn voor gebruik door hulpverleners die gespecialiseerd zijn in angst- en psychosomatische stoornissen.

Het vinden van aanvullende factoren voor de originele structuur lijkt erop te wijzen dat er geen conceptuele equivalentie is tussen de twee versies van de OQ. Ook was de *fit* van onze drie-factoroplossing beduidend beter dan die van de originele drie-factoroplossing. Dit houdt echter niet noodzakelijkerwijs in dat ze niet equivalent zijn. Enkele van de *goodness-of-fit* indices die we voor de drie-factoroplossing gevonden hebben, kwamen overeen met de indices voor de Amerikaanse steekproef (Mueller et al, 1998). Als onze statistische analyses toegepast zouden worden op de Amerikaanse gegevens, wordt mogelijk een vergelijkbare structuur voor de Amerikaanse OQ gevonden.

Een vergelijking tussen de normatieve scores van de Amerikaanse en de Nederlandse bevolking heeft aangetoond dat de Nederlandse steekproeven van de normale populatie en de Nederlandse klinische steekproeven iets lager scoorden dan hun Amerikaanse equivalenten. Het is niet ongebruikelijk voor psychologische tests dat er verschillen zijn tussen de scores van verschillende bevolkingen. Hoewel de verschillen relatief klein waren, moesten de *cut-off* scores gekalibreerd worden om de criteriumequivalentie te verbeteren. Als gevolg van de kalibratie ligt de *cut-off* score voor de Nederlandse bevolking 8 punten lager dan de Amerikaanse *cut-off* score. Na kalibratie waren de waarden van de sensitiviteit en de specificiteit vergelijkbaar met

de waarden van de originele versie. De specificiteit van de Amerikaanse en de Nederlandse OQ is respectievelijk 0.83 en 0.85 en de sensitiviteit is 0.84 voor beide versies.

Een opvallend verschil tussen de Amerikaanse en Nederlandse normatieve scores is dat er bij de Nederlandse bevolking sekseverschillen gevonden werden bij zowel de klinische steekproef als de steekproef onder de normale populatie. Mannen hadden meer problemen op het gebied van de Sociale Rol, terwijl vrouwen hogere niveaus van Symptomatische Distress en van Angst en Somatische Distress vertoonden. Sekseverschillen in normatieve scores zijn vrij gebruikelijk bij dit soort tests en het is eerder verrassend dat er geen verschillen gevonden werden in de originele OQ. In de handleiding staat vermeld dat er bij één onderzoek enkele sekseverschillen werden gevonden onder patiënten, maar dat deze niet klinisch relevant bevonden werden. Er wordt echter maar over kleine aantallen per sekse gerapporteerd, dus een gebrek aan *power* kan de oorzaak zijn van de niet-significante resultaten.

Behalve de equivalentie, werden ook de psychometrische kenmerken van de Nederlandse versie onderzocht. De betrouwbaarheid van de subschalen en van de totale schaal was bij de meeste steekproeven toereikend. Een uitzondering was de interne consistentie van het domein van de Sociale Rol, die te laag was voor alle drie de steekproeven. De gevoeligheid voor verandering is heel goed en de OQ kan goed onderscheid maken tussen functionele en disfunctionele populaties. De concurrente validiteit is goed voor de SD en ASD subschalen, maar minder goed voor de IR en SR subschalen. De lagere correlaties met de IR en SR subschalen waren echter te verwachten, gezien de conceptuele verschillen tussen de GVSG-45 en de OQ subschalen. Dit betekent niet noodzakelijkerwijs een lage validiteit voor deze schalen.

Dit onderzoek heeft zich niet beziggehouden met de OQ als meetinstrument om de vooruitgang van de patiënten te volgen. Er moet meer onderzoek met de Nederlandse OQ verricht worden op dit terrein om een betere vergelijking te kunnen maken tussen de vooruitgangscurven van de Nederlandse en Amerikaanse populaties. Gezien de gevonden verschillen in normatieve scores, zijn ook verschillen in behandelverloop te verwachten. Er moet ook meer onderzoek verricht worden naar het domein van de Sociale Rol en de aanvullende factor die werd gevonden in dit onderzoek.

De Nederlandse OQ heeft vergelijkbare psychometrische kenmerken als het originele instrument, maar de twee versies zijn niet equivalent in alle aspecten. Er is mogelijk een verschil in conceptuele equivalentie, hoewel nadere analyses van het originele instrument het tegendeel kunnen bewijzen. De criteriumvaliditeit van de Nederlandse OQ lag dicht bij de originele waarden, maar pas na de kalibratie van de grensscores, wat een gebrek aan criteriumequivalentie aangeeft. Deze resultaten laten zien dat gelijksoortige psychometrische kenmerken geen equivalentie garanderen. Het feit dat het nodig was om de *cut-off* scores te kalibreren, ondanks kleine verschillen tussen de scores van de twee bevolkingen, toont aan hoe belangrijk het is om goede normatieve scores te hebben voor vertaalde instrumenten. Dit is vooral het geval voor klinische uitkomstinstrumenten zoals de OQ, waar het gebruik van inadequate normen zou kunnen leiden tot verkeerde behandelbeslissingen.

Noten

1. In het oorspronkelijke artikel worden globale en specieke instrumenten gerapporteerd. In dit artikel is ervoor gekozen alleen de globale instrumenten weer te geven.

Literatuur

Arrindell, W. A., & Ettema, J.H.M. (1975). *Klachtenlijst (SCL-90)*. Lisse: Swets & Zeitlinger.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Cronbach, L. J. (1990). *Essentials of psychological testing*. New York, NY: Harper Collins Publishers.

Beurs, E. de, Dyck, R. van, Marquenie, L.A., Lange, A., & Blonk, R.W.B. (2001). De DASS: een vragenlijst voor het meten van depressie, angst en stress. *Gedragstherapie, 34*, 35-53.

Jong, A. de, & Lubbe, P.M. van der (2001). *Groningse vragenlijst voor Sociaal Gedrag: Zelfbeoordelingsvrageniljsten voor het vaststellen van problemen in het interpersoonlijk functioneren (handleiding)*. Groningen: Rob Giel Onderzoekscentrum/Rijksuniversiteit Groningen, disciplinegroep Psychiatrie.

Jong, K. de, & Nugter, M.A. (2004). De Outcome Questionnaire: psychometrische kenmerken van de Nederlandse vertaling. *Nederlandsch Tijdschrift voor Psychologie, 59*, 76-79.

Derogatis, L.R. (1975). *The Brief Symptom Inventory*. Baltimore, MD: Clinical psychometrics unit.

Derogatis, L.R. (1977). *The SCL-90 Manual: Scoring, admininstration and procedures for the SCL-90*. Baltimore, MD: John Hopkins University School of Medicine, Clinical psychometrics unit.

Flaherty, J.A., Gaviria, F.M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J.A. & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *Journal of Nervous Mental Disorders, 176*, 257-63.

Gladis, M., Gosch, E.A., Dishuk, N.M., & Crits-Christoph, P. (1999). Quality of life: expanding the scope of clinical significance. *Journal of Consulting & Clinical Psychology, 67*, 320-331.

Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Kline, R.B. (1998). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

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

Lambert, M.J., Hannöver, W., Nisslmüller, K., Richard, M., & Kordy, H. (2002). Fragebogen zum ergebnis von psychotherapie: Zur reliabilität und validitat der deutschen ubersetzung des Outcome Questionnaire 45.2 (OQ-45.2). *Zeitschrift fur Klinische Psychologie und Psychotherapie: Forschung und Praxis, 31*, 40-46.

Lambert, M.J., Morton, J.J., Hatfield, D.R., Harmon, C., Hamilton, S., Shimokawa, K., et al. (2004). *Administration and scoring manual for the OQ-45.2 (Outcome Questionnaire)* (3 ed.). Wilmington, DE: American Professional Credentialling Services LLC.

Lambert, M.J. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17*, 1-14.

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

Thalheimer, W., & Cook, S. (2002). *How to calculate effect sizes from published research articles: A simplified methodology*. Retrieved February, 2006, from http://work-learning.com/effect_sizes.html

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