Clinical Psychology and Psychotherapy
Clin. Psychol. Psychother. **14**, 288–301 (2007)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/cpp.529

Assessment

The Outcome Questionnaire (OQ-45) in a Dutch Population: A Cross-Cultural Validation

Kim de Jong,[1]* M. Annet Nugter,[1] Marike G. Polak,[2] Johannes E. A. Wagenborg,[3] Philip Spinhoven[2,4] and Willem J. Heiser[2]

[1] *Research Department, GGZ Noord-Holland-Noord, Heiloo, the Netherlands*
[2] *Department of Psychology, Leiden University, the Netherlands*
[3] *Department of Research and Development, de Geestgronden, Bennebroek, the Netherlands*
[4] *Department of Psychiatry, Leiden University, the Netherlands*

The cross-cultural validity of the Outcome Questionnaire (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 from a university (*n* = 268), in a community (*n* = 810) and from three mental health care organizations (*n* = 1920). Results show 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 a clinically significant change in the Dutch population. Copyright © 2007 John Wiley & Sons, Ltd.

INTRODUCTION

Over the last years, the Outcome Questionnaire (OQ; Lambert et al., 1996) has become one of the 10 instruments most frequently used by practitioners in the USA to measure clinical outcomes (Hatfield & Ogles, 2004) and is often used in clinical outcome research. It is also gaining popularity in other countries and has been translated into several languages, including Japanese, Korean, Italian, French, Portuguese, German and Dutch. Even though the psychometric properties of the original version of the OQ have been thoroughly investigated, few papers are available on the properties of translated versions of the OQ. This paper addresses the cross-cultural validity of the Dutch OQ.

Reasons for the popularity of the OQ lie in the fact that it has some advantages that most other

*Correspondence to: Kim de Jong, MA, GGZ Noord-Holland-Noord, Research Department, Postbus 18, 1850 BA Heiloo, the Netherlands.
E-mail: k.dejong@ggz-nhn.nl

instruments do not have. First, the OQ aims to measure three domains of functioning: symptom distress (SD), interpersonal relations (IR) and social role (SR) performance. It has become accepted in outcome research to measure symptom reduction as well as an improvement in well-being (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999). Other popular outcome instruments such as the Symptom Checklist-90 (SCL-90; Derogatis, 1977), the Brief Symptom Inventory (Derogatis, 1975) or the Social Adjustment Scale (SAS; Weissman & Bothwell, 1976) measure either symptoms or functioning. Additional instruments are required to measure the other domains. Moreover, the OQ, along with instruments such as the Clinical Outcomes in Routine Evaluation–Outcome Measure (Evans et al., 2000), is a general instrument that can be used for a large variety of disorders, so a comparison in functioning and outcome of a broad range of patients is possible, irrespective of psychiatric diagnosis. By contrast, specific instruments are designed for certain disorders and cannot compare symptoms and functioning of different types of patients.

Also, the OQ is relatively short. An average patient can complete it in approximately 5 minutes. This is especially important in clinical practice, where there is neither time nor budget for the test batteries that are common in academic research.

Probably the most important feature of the OQ is that it is capable of tracking patient progress by repeated measurements. The OQ is frequently used in outcome research that provides weekly feedback to the therapist about the patient's progress. The patient's treatment course is compared to a predicted course, and the therapist is alerted when the patient goes too far off the predicted track. This feedback results in more effective treatments, especially for those patients who have a higher risk for treatment failure (Lambert, Harmon, Slade, Whipple, & Hawkins, 2005; Lambert et al., 2001, 2003).

Psychometric Properties

The psychometric properties of the American version of the OQ have been extensively investigated. Reliability and validity estimates are good for the SD subscale and total scale. Reliability is adequate for the IR and SR subscales, but concurrent validity estimates are less convincing. The OQ has proper sensitivity and specificity, and the sensitivity to change is good on the scale as well as on the item level (Lambert et al., 2004). Normative scores of the clinical and community samples differ significantly. The structure of the instrument, with three subscales, does not seem to have sufficient empirical support. Mueller, Lambert, and Burlingame (1998) found that the three-factor model did not have a proper fit in a confirmatory factor analysis. Chapman (2003) could not find support for the three-factor structure either and performed an exploratory analysis, which resulted in nine factors. However, these factors have not been confirmed in a new sample.

Equivalence between Language Versions

During translation, many problems may occur that change the properties of an instrument. First, there may be a difference in the meaning of the items of the instrument in a different language. Flaherty et al. (1988) refer to this as *semantic equivalence* and state that the key to obtain semantic equivalence is the back-translation method, which has been used in all OQ translations. Still, a slight semantic difference is likely to occur, as the English language has considerably more words than the Dutch.

In order to assure the equivalence of two language versions of psychological tests, the constructs underlying the test need to be equivalent (Butcher, Derksen, Sloore, & Sirigatti, 2003). Cultural differences influence the *conceptual equivalence* of two language versions of an instrument. Hofstede (2006) introduced five cultural dimensions that can assist in differentiating between cultures. His research showed that American and Dutch cultures have similar levels of individualism, power distance and uncertainty avoidance, but differ in long-term orientation and masculinity.

One method of examining conceptual equivalence is to determine whether the items and scales maintain generally the same factors in the new language version. A well-known example of a difference in conceptual equivalence is the SCL-90: the Dutch version has a different factor structure than the original version (Arrindell & Ettema, 1975).

Differences in normative scores often exist between different cultures, even among Western countries. For example, in the European Psychiatric Services: Inputs Linked to Outcomes and Needs (EPSILON) study, van Wijngaarden et al. (2000) found differences in scoring and reliability estimates on the Involvement Evaluation Questionnaire: scores were usually high in Verona and low in Copenhagen. The Minnesota Multiphasic

DOI: 10.1002/cpp

Personality Inventory (MMPI) also has different norms for different countries (Butcher et al., 2003). More importantly, a difference in normative scores may result in different criterion validity, also changing the sensitivity and specificity of the instrument. Flaherty et al. (1988) refer to this as a lack of *criterion equivalence*.

In clinical outcome research, a common criterion is whether clinical significant change (Jacobson & Truax, 1991) has occurred between post- and pretreatment measurements. Because the cut-off point for estimating clinical significant change is based on population curves, it is important to take proper samples of these populations in the new culture. If criterion equivalence is not reached, a calibration of scores should be performed, thus calculating new cut-off scores for the population (Flaherty et al., 1988). Differences in cut-off scores are usually the result of significant differences in normative mean scores between two cultures, but may also occur when differences in mean scores are small (non-significant). Furthermore, differences in reliability estimates can lead to different reliable change indices (RCIs), because in the calculation of the RCI, the reliability of the instrument is used to estimate the measurement error.

The goal of the current research project was to find out if the factor structure and normative scores are equivalent with the original version and to determine the psychometric properties of the Dutch OQ. Preliminary studies showed that Dutch subjects seemed to have lower scores on the OQ than American subjects (de Jong & Nugter, 2004). Also, some differences in psychometric properties were found. At that time, however, sample sizes were not large and representative enough to draw definite conclusions. Because the normative sample size was insufficient for individual use of the OQ and more information was needed on the validity, more data were collected and additional studies were performed. For this paper, the data of initial and further studies of the OQ are combined.

METHOD

The Data

Subjects included a university sample, two community samples and two clinical samples. All data were collected for research purposes only. A student sample of 268 undergraduates was collected from the psychology department of the University of Amsterdam as part of its educational program. Subjects completed an OQ, SCL-90 and Groningse Vragenlijst Sociaal Gedrag-45 (GVSG-45) at the first session. A group of 264 students completed the OQ for the second time after 2 weeks. Data from the student sample were used only to calculate reliability and validity, as this group was not representative to function as a normative group.

A community sample of 446 individuals was collected by a random selection of subjects in the phone directory of 13 phone regions, stratified by province, geographically distributed through the Netherlands and including major cities as well as suburban and rural regions. All adults in the household were asked to complete the questionnaire. When they consented to participate, questionnaires and consent forms were mailed. A selection of 1270 numbers was made of which 286 were never reached and 487 households consented. A total of 818 questionnaires were sent (with an average of 1.7 per household), and 446 were returned completed (55%).

Another community sample of 362 individuals was collected from 14 commercial and non-profit business settings in a variety of business branches. The questionnaires were distributed by internal mail. A stamped addressed envelope was included. Completion of the test was on a voluntary basis and was anonymous. A total of 1097 questionnaires was spread at the business settings, so the response rate was 33%, which is average for this type of sample.

Subjects in the community sample who received treatment for psychological or psychiatric problems were deleted from the sample; 24 subjects were removed on this ground.

The first clinical outpatient sample of 1545 persons was collected at four sites of three public mental health care institutions. Patients completed a paper-and-pencil version of the OQ either before or after intake. To obtain test–retest stability data, a subsample of 43 patients who completed the OQ after the intake completed the OQ 2–3 weeks later, before they received a treatment advice. During that time period, no treatment sessions took place. A different subsample of 117 patients completed the SCL-90 and Depression Anxiety and Stress Scale (DASS) together with the OQ after intake to obtain concurrent validity estimates.

By means of an internet screening tool, an additional sample of 375 patients completed the OQ online prior to their intake session. Depending on a global screening procedure, additional specific questionnaires that matched the subject's symptoms were completed by the subjects. An average

DOI: 10.1002/cpp

Table 1. Characteristics of the samples

Sample	*n*	Gender		Age	
		Female *n* (%)	Male *n* (%)	Range	Mean (SD)
Community sample	810	513 (63)	297 (37)	18–94	44.3 (15)
Phone directory sample	446	247 (55)	199 (45)	18–94	48.1 (16)
Business sample	361	264 (73)	97 (27)	18–77	39.5 (12)
University sample	268	171 (64)	96 (36)	17–53	22.3 (6)
Clinical sample					
Outpatient sample paper-and-pencil	1545	896 (58)	628 (41)	18–65	37.3 (11)
Test–retest	42	31 (74)	11 (26)	18–55	31.7 (10)
Concurrent validity SCL-90/DASS	118	76 (64)	41 (35)	18–61	33.6 (11)
Sensitivity to change	60	32 (53)	24 (40)	23–62	41.6 (10)
Outpatient internet screening tool	375	–	–	–	–

SD = Standard Deviation. SCL-90 = Symptom Checklist-90. DASS = Depression Anxiety and Stress Scale.

number of 9.6 (Standard Deviation [SD] = 3.3) questionnaires were completed by the patient.

An overview of sample characteristics is given in Table 1. There are some differences between the community and business samples with regard to gender and age, because one of the companies in the business sample, a private home care organization, had relatively many females among their employees. Data from this company were kept in the sample because scores on the OQ did not differ significantly from the other business sites. For technical reasons, demographic characteristics of the internet screening sample were not available.

Subjects who left more than 20% of the questions of the questionnaire unanswered were removed from the sample; 8 students, 4 persons from the community sample and 49 outpatients were deleted. In case of missing values, a mean score for the remaining scale items was calculated, multiplied by the number of items on the scale and rounded to the nearest number. Mean subscale scores were only calculated if no more than 20% of the scale items were missing. Missing values were not replaced in those analyses that make use of the data on the item level (e.g., factor analysis, reliability analysis).

Instruments

The OQ

The OQ consisted of 45 items that were scored on a five-point rating scale, ranging from never (0) to almost always (4). The SD subscale had 25 items that were associated with most common disorders in public mental health care; depression, anxiety and addiction to alcohol or drugs were well represented. The IR subscale consisted of 11 items and measured the functioning of the patient in relationships with partner, family and friends. The SR subscale contained nine items and measured functioning in school, work and leisure. There were nine reversely scored items.

The American normative sample consisted of undergraduate, community and clinical subsamples. The undergraduate samples were collected from universities in three states; the community sample was collected from various business locations and from the Utah phone directory, and the clinical samples were collected from a university counselling centre, an employee assistance program, a university-based outpatient clinic and a community mental health service centre (Lambert et al., 2004). Comparisons with the Dutch samples were made with the undergraduate, community and outpatient samples.

Instruments Used for Validation of the OQ

A short description of the instruments used for validation of the OQ is given in this section. All instruments are self-rating questionnaires and have proper psychometric properties. References for both the original and the Dutch versions are given. A distinction is made between general and specific instruments.

General Instruments. The SCL-90-item version (Arrindell & Ettema, 1975) and DASS (de Beurs, van Dyck, Marquenie, Lange, & Blonk, 2001; Lovibond & Lovibond, 1995) were used to validate the SD subscale. For the SCL-90, the Global Severity Index (GSI) was calculated. For the DASS, subscale scores were used to correlate with the OQ subscales.

DOI: 10.1002/cpp

The GVSG-45-item version (de Jong & van der Lubbe, 2001) was used to validate the IR and SR domain scores of the OQ. It measures social behaviour on nine domains. For this research, two indices that were not in the original questionnaire were calculated: as an index of interpersonal problems, we used the mean score of the Parents, Partner, both Children domains and the Friends domain, further referred to as Functioning on Interpersonal Relationships. As an index of SR performance, we calculated the mean score of the School, Occupation, Housework and Leisure domains, further referred to as Functioning on Social Role.

Specific Instruments. The OQ claims to be usable for a variety of disorders. Therefore, a number of instruments that aim to measure symptoms of specific disorders were compared with the OQ. The specific instruments were part of an internet screening tool. Not all instruments that were in the internet screening tool were used in analysis. The instruments were selected using two criteria. First, the number of patients who completed the instrument had to exceed 30. Second, the instruments had to measure symptoms that occur in a variety of patients, such as anxiety, depression, grief and reactions to overwhelming experiences. One would expect medium-high correlations between the specific instruments and the SD subscale of the OQ and low-to-medium correlations with the IR and SR subscales.

The Quick Inventory of Depressive Symptoms Self-Report (QIDS-SR16; Rush et al., 2003) assesses all the criterion symptoms that the *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition* indicates to diagnose a major depressive episode.

The Body Sensations Questionnaire (BSQ) and Agoraphobic Cognitions Questionnaire (ACQ) (Bouman, 1995; Bouman, 1998; Chambless, Caputo, Bright, & Gallager, 1984) both measure experienced anxiety during panic or anxiety attacks. The BSQ measures anxiety for physical sensations, while the ACQ measures catastrophic cognitions during the anxiety or panic attack.

The Liebowitz Social Anxiety Scale-Self-Report (LSAS-SR; Liebowitz, 1987; van Balkom, de Beurs, Hovens, & van Vliet, 2004) measures the severity of social phobic symptoms and avoidance behaviour.

The Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990; van Rijsoort, Vervaeke, & Emmelkamp, 1999) measures the inclination to worry as well as the amount, intensity and uncontrollability of worrying.

The Padua Inventory-Revised (Sanavio, 1988; Van Oppen, Hoekstra, & Emmelkamp, 1995) measures obsessive thoughts and compulsions on five domains: impulse, wash, control, ruminate and precision. The total score is used to assess the severity of obsessive-compulsive symptoms.

The Impact of Events Scale Revised (IESR; Weiss & Marmar, 1997) consists of 22 items measuring re-experience of shocking events and avoidance of thoughts and feelings that are related to these events. It also measures increased arousal. There are three subscales: Intrusion, Avoidance and Hyperarousal. The Hyperarousal subscale was not yet normed for the Dutch population and is therefore not included in the total score. The Dutch translation of the IESR is known as the Schokverwerkingslijst (Brom & Kleber, 1985; van der Ploeg, Mooren, Kleber, van der Velden, & Brom, 2004).

The Inventory of Complicated Grief-revised (ICG-r; Prigerson, Kasl, & Jacobs, 1997) measures symptoms of normal and potentially complicated grief and mourning. The Dutch translation of the ICG-r is known as the Rouw Vragenlijst (Boelen, de Keijser, & van den Bout, 2001).

Analysis

The paper-and-pencil data were scanned by computer and validated using the Teleform software from Verity, Inc.; Sunnydale, CA; USA, version 9.0. Data from the internet screening tool were collected with the screening tool developed specifically for one of the mental health care organizations by Interapy (2004).

All tests of difference were two-tailed against $p < 0.05$. The large sample sizes gave high statistical power; therefore, effect sizes were also reported. Effect sizes (Cohen's d) were calculated with the effect size spreadsheet by Thalheimer and Cook (2002) in Microsoft Excel and were interpreted according to the criteria reported by Cohen (1992). Most analyses were conducted in SPSS for Windows, version 13.0 (2004) SPSS Inc. Chicago, IL: USA. The confirmatory factor analysis was performed in EQS 6.1 (2005). Multivariate Software, Inc. Encino, CA: USA. Exploratory principal component analysis on the residual matrix was performed in MATLAB, release 14 (2005) The Mathworks, Inc. Natick, MA: USA.

The goodness-of-fit indices that were reported in the confirmatory factor analysis were the root mean square residual (RMR), the root mean square error of approximation (RMSEA), the GFI, the Bentler–Bonnet normed fit index (NFI), the com-

Table 2. Confirmatory factor analysis GFIs

Model	χ^2	df	χ^2/df	RMR	RMSEA	GFI	NFI	CFI
Three-factor solution	3678.2	942	3.90	0.103	0.046*	0.880	0.933*	0.949*
Five-factor solution	3413.4	925	3.69	0.075*	0.044*	0.889	0.957*	0.964*

*Meets the recommended criteria.
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. χ^2 = chi-square. df = degrees of freedom.

parative fit index (CFI), the chi-square (χ^2) and the chi-square divided by degrees of freedom in the model (χ^2/df). General guidelines were that the RMR should be less than 0.10, the RMSEA less than 0.05, the GFI greater than 0.95, the CFI and NFI greater than 0.90, a χ^2 that is non-significant and χ^2/df less than 2 (Kline, 1998). The goodness-of-fit indices that were reported in the original study by Mueller et al. (1998) are also reported in the current paper, except for the adjusted goodness of fit index and critical N. The RMSEA was added.

RESULTS

Equivalence

Conceptual Equivalence

Factor Analysis. As was stated earlier, the factor structure of an instrument may change in a different cultural setting or language. The sample, consisting of the community and clinical sample, was randomly split in two so that if additional analyses would result in new factor structures, they could be fitted on the other split of the sample to test the stability of the factor solution. The first sample was used to test whether the OQ had a three-factor structure. A confirmatory factor analysis was conducted, using general least squares estimation to make our solution comparable to the original analyses of Mueller et al. (1998).[1]

As can be seen in Table 2, our solution meets three out of seven goodness-of-fit criteria. The indices that did not meet the criteria were the χ^2, χ^2/df, RMR and GFI. The χ^2 and χ^2/df are dependent on the sample size, and given our large sample size and consequently high power, a significant χ^2 is not necessarily a sign of a poor fit. The criteria for a good fit for the RMSEA, NFI and CFI criteria are met and the three-factor solution seems to have a reasonable fit. The fit of our solution is notably better than the solution that Mueller et al. (1998) obtained; their three-factor solution met none of the criteria that we applied.

Table 3 shows the standardized factor loadings on the three factors. Four items have factor loadings below 0.30: items 11, 14, 26 and 32. Items 11, 26 and 32 are known difficult items. They all measure problematic alcohol/drug use and have rather skewed scoring (a lot of '0' scores). In exploratory factor analysis, they consequently end up in one factor, which consists only of these three items. Item 14 'I work/study too much' is another special case. This item does not perform well in the original OQ also and shows a negative correlation with several items in the covariance matrix. Moreover, it is the only item in the SR subscale wherein the functional sample (M = 1.87, SD = 1.1) actually scores slightly higher than the clinical sample (M = 1.68, SD = 1.2), $t(1706) = 3.54$, $p < 0.001$, $d = 0.14$.

As the three-factor solution still was not satisfactory, we tried to explain more variance by using the residual matrix from the three-factor solution. A principal component analysis with varimax rotation produced two components with an eigenvalue greater than 1, which explained 34% of variance. Because of the varimax rotation, the two components are not correlated, and as the solution was based on the residuals of the three-factor solution, they are considered independent from the first three factors as well. For each component, the items that loaded above 0.15 were selected.

The two components were added to the original three-factor model and fitted on the second split sample. The chi-square value of the five-factor solution has dropped 264.7 points, associated with a loss of 17 degrees of freedom, which leads to a χ^2/df improvement ratio of 15.6. A ratio of 2 is usually considered a substantial improvement. The other goodness-of-fit values also show an improvement of fit. Especially important is the RMR value, which did not meet the criterion of 0.10 in the

[1] Maximum likelihood is the standard estimation technique in confirmatory factor analysis.

DOI: 10.1002/cpp

Table 3. Standardized factor loadings of the factor models

Item	Three-factor solution (n = 1362)			Five-factor solution (n = 1363)				
	F1	F2	F3	F1	F2	F3	F4	F5
2	0.66			0.65				0.16
3	0.63			0.62				
5	0.71			0.73				
6	0.73			0.72				
8	0.67			0.66				
9	0.85			0.82				0.13
10	0.78			0.67				0.38
11	0.18			0.14				
13	0.84			0.87				
15	0.84			0.87				
22	0.68			0.68				
23	0.80			0.83				
24	0.75			0.77				
25	0.76			0.75				0.14
27	0.45			0.36				0.34
29	0.61			0.51				0.44
31	0.86			0.88				
33	0.71			0.61				0.32
34	0.37			0.33				0.39
35	0.48			0.38				0.39
36	0.73			0.68				0.29
40	0.73			0.71				
41	0.57			0.56				0.25
42	0.84			0.87				
45	0.45			0.41				0.34
1		0.59			0.60			
7		0.60			0.60			
16		0.45			0.41			0.21
17		0.46			0.44			
18		0.81			0.81			
19		0.58			0.62			
20		0.71			0.76			
26		0.27			0.27			
30		0.74			0.73			
37		0.70			0.71			
43		0.80			0.84			
4			0.71			0.51	0.39	
12			0.63			0.68		
14			0.13			–0.11	0.34	
21			0.76			0.73		
28			0.34			0.30		
32			0.21			0.20		
38			0.81			0.66	0.50	
39			0.65			0.44	0.48	
44			0.66			0.58		

three-factor solution, but does so in the five-factor solution.

The first additional factor consists of items that all belong to the SR domain. The unexplained variance in the SR domain in the three-factor solution seems to be mainly caused by item 14, which has a low factor loading on this domain. In the five-factor solution, this item even has a negative factor loading on the SR domain. For clinical purposes, the extra social factor does not add much. It may be useful if one is merely interested in the social role functioning of an individual, but in that case, the OQ would probably not be used as the instrument of choice.

In contrast, the second additional factor does seem to add something to the clinical utility of the instrument. Most of the items on this factor originate from the SD scale, which is a rather long scale. The items seem to be related to anxiety and somatic manifestations of anxiety. Some of the items represent cognitive representations of anxiety, such as item 10, 'I feel fearful', whereas others seem more physical manifestations of anxiety, such as item 29, 'My heart pounds too much', which is known to be a symptom of anxiety or panic attack. It may be a useful addition to the original three-factor structure. Therefore, we decided to evaluate the validity of this factor, along with the validity of the original three factors. The factor is further referred to as Anxiety and Somatic Distress (ASD).

Correlations between Subscales. The correlations between the subscales of an instrument give an indication of whether the structure of the instrument is as it was intended. In the case of an instrument that assesses several domains of functioning, multidimensionality should be reflected in the factor structure of the instrument. Also, each subscale should assess a concept that is not measured by the other subscales. Therefore, the correlations between the domains should not be too high. Table 4 shows that the correlation between the subscales of the OQ is higher than is desirable, indicating a moderate construct validity. Especially high is the correlation between the SD and ASD subscales. This is not surprising, as the ASD subscale consists almost exclusively of items that are in the SD scale, but considering that, correlations would ideally be lower.

Criterion Equivalence

Differences in Scoring

Table 5 shows the mean scores of Dutch and American samples on the OQ subscales and total scale. The mean scores for the Dutch community ($t = 7.48$, $p < 0.001$, $d = 0.37$) and clinical samples ($t = 2.50$, $p = 0.01$, $d = 0.15$) are somewhat below the

DOI: 10.1002/cpp

Table 4. Correlations between the subscales and total scale

	SD	ASD	IR	SR	Total
Symptom Distress (SD)	1.0 (n = 2726)				
Anxiety and Somatic Distress (ASD)	0.94 (n = 2726)	1.0 (n = 2726)			
Interpersonal Relations (IR)	0.75 (n = 2723)	0.64 (n = 2723)	1.0		
Social Role (SR)	0.68 (n = 2646)	0.60 (n = 2646)	0.59 (n = 2644)	1.0	
OQ total score	0.97 (n = 2726)	0.89 (n = 2726)	0.85 (n = 2724)	0.77 (n = 2647)	1.0 (n = 2727)

OQ = Outcome Questionnaire.

Table 5. Means and Standard Deviations (SDs) of OQ in the Dutch and American samples

	American samples						Dutch samples					
	University (n = 235)		Community (n = 815)		Clinical (n = 342)		University (n = 268)		Community (n = 807)		Clinical (n = 1920)	
	M	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD
Symptom Distress	23.0	10	25.4	12	49.4	15	27.3	12	22.2	10	48.9[†]	16
Anxiety and Somatic Distress	–	–	–	–	–	–	15.6	7	13.3	6	25.9[†]	9
Interpersonal Relations	8.8	5	10.2	6	19.7	6	11.4	5	8.4	5	16.8[‡]	7
Social Role	10.4	4	9.6	4	14.1	5	10.4	4	8.1*	3	13.6[§]	6
Total score	42.2	17	45.2	19	83.1	22	49.1	18	38.7	16	79.5	25

Notes: Means and SDs of the American samples were copied from the OQ manual (Lambert et al., 2003).
*798 cases.
[†] 1919 cases.
[‡] 1917 cases.
[§] 1849 cases.
OQ = Outcome Questionnaire.

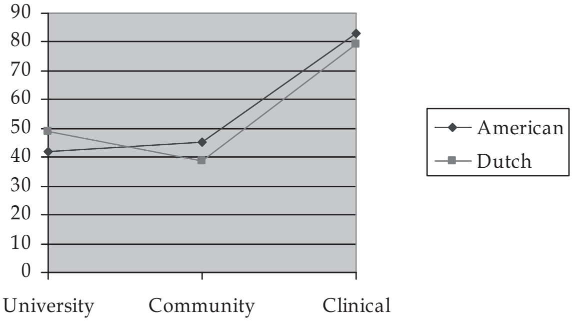

Figure 1. Outcome Questionnaire total score for the American and Dutch samples

American equivalents, even though the effect sizes are small. The mean scores of Dutch students are somewhat higher than the American student sample ($t = -4.40$, $p < 0.001$, $d = 0.39$). Figure 1 gives a visual representation of the sample differences for the OQ total scale.

In the American samples, no differences were found between males and females. In the Dutch samples, some small differences were found. In Table 6, scores for the clinical and community samples are given for males and females. In the community sample, significant differences were found for gender, Wilks' $\lambda = 0.92$, $F(5, 792) = 13.3$, $p < 0.001$. Women showed higher levels of SD, ($F[1, 796] = 10.7$, $p = 0.001$, $d = 0.26$) and ASD ($F[1, 796] = 21.8$, $p < 0.001$, $d = 0.37$), while men showed more problems in SR performance ($F[1, 796] = 13.6$, $p < 0.001$, $d = 0.27$).

Similar results were found in the clinical sample: Wilks' $\lambda = 0.92$, $F(5, 1446) = 23.7$, $p < 0.001$. Here, women showed slightly higher levels of SD ($F[1, 1450] = 5.83$, $p = 0.016$, $d = 0.13$) and ASD ($F[1, 1450) = 29.1$, $p < 0.001$, $d = 0.29$), while men showed some-

DOI: 10.1002/cpp

Table 6. Means and Standard Deviations (SDs) by gender in the clinical and non-clinical samples

	Community sample						Clinical sample					
	Male			Female			Male			Female		
	n	*M*	SD	*n*	*M*	SD	*n*	*M*	SD	*n*	*M*	SD
Symptom Distress	296	20.6	10	511	23.2	10	628	47.4	15	896	49.3	16
Anxiety and Somatic Distress	296	11.9	6	511	14.1	6	628	24.3	9	896	26.8	9
Interpersonal Relations	296	8.3	5	511	8.4	5	627	16.6	7	894	16.6	7
Social Role	292	8.6	4	506	7.7	3	598	14.3	5	857	12.8	5
OQ total score	296	37.4	16	511	39.4	16	628	78.4	25	896	79.0	25

OQ = Outcome Questionnaire.

what more problems in SR performance ($F[1, 796] = 25.4$, $p < 0.001$, $d = 0.27$).

Clinical Significance and Reliable Change

To measure individual change, the criterion of clinical significance by Jacobson and colleagues is often applied (Jacobson & Truax, 1991; Jacobson, Follette, & Revenstorf, 1984; Jacobson, Roberts, Berns, & McGlinchey, 1999). Their criterion is twofold: (1) the magnitude of change has to be statistically reliable, and (2) by the end of treatment, patients have to end up in a (score) range that renders them indistinguishable from well-functioning people. A cut-off point for clinically significant change and an RCI are calculated using formula c by Jacobson and Truax (1991).

For the American OQ, the cut-off score for clinical dysfunctioning is 63 for the total scale and 36, 15 and 12, respectively, for the SD, IR and SR subscales. The RCI is 14 for the total scale and 10, 8 and 7 for the subscales, respectively (Lambert et al., 2004). For the Dutch OQ, the cut-off score for the SD subscale is 33; for the ASD subscale, it is 19; for the IR subscale, it is 12; for the SR subscale, it is 10, and for the total scale, the cut-off score is 55. A person that scores on or above the cut-off score belongs to the dysfunctional (clinical) range. Given the differences found between male and female respondents, separate cut-off scores for men and women were also calculated for the subscales with gender differences. The cut-off score for the SD subscale was 31 for men and 33 for women; the cut-off for the ASD subscale was 17 for men and 19 for women, and the cut-off for the SR subscale was 12 for men and 10 for women.

Using the cut-off score of 55, sensitivity for the OQ total scale is 0.84, which means that 84% of the community sample is correctly identified as belonging to the functional sample. The specificity of the OQ is 0.85, which means that 85% of the clinical sample is correctly identified as dysfunctional. Using the cut-off score of 63 from the original OQ leads to a higher sensitivity (0.93), but lower specificity (0.74). This means that fewer patients are correctly identified as belonging to the dysfunctional sample.

The RCI is usually expressed as the amount of points on a certain measurement instrument that a patient has to improve between pre- and post-treatment measurements. The RCI depends on the reliability of the measurement instrument and the variability of scores. As reliability index, the pooled internal consistency of the clinical and community sample was used (see Table 7). Using either subgroup would lead to less variability in the answers, which leads to lower values of Cronbach's alpha. In literature, this phenomenon is referred to as range restriction (Cronbach, 1990). The RCIs for the SD, ASD, IR and SR subscales are 10, 8, 8 and 9, respectively. The RCI for the OQ total scale is 14, so a patient has to improve a minimum of 14 points on the OQ to obtain reliable change.

Psychometric Properties

Reliability

Internal consistency estimates are sufficient for subscales and the total scale in most of the samples (see Table 7), except for the SR subscale, for which disappointing values for Cronbach's alpha were found in the university, community and clinical samples. Combining the clinical and community samples improves the results, which indicates that restriction of range may occur here. Another explanation may be that an increased sample size improves internal consistency values. In reliability analysis, cases are rapidly lost: if one item of the scale is missing, the case cannot be used entirely.

DOI: 10.1002/cpp

Table 7. Internal consistency (Cronbach's alpha [*a*]) est–retest reliability (Pearson's product-moment correlation coefficient)

Domain	Internal consistency						Test–retest			
	University		Community		Clinical		Community and clinical		University (n = 264)	Clinical (n = 42)
	n	a	n	a	n	a	n	a	r	r
Symptom Distress	257	0.90	768	0.89	1247	0.91	2390	0.95	0.81	0.76
Anxiety and Somatic Distress	261	0.79	786	0.82	1743	0.84	2529	0.89	0.74	0.70
Interpersonal Relations	264	0.74	770	0.77	1607	0.80	2377	0.84	0.71*	0.83
Social Role	258	0.61	773	0.53	1620	0.69	2393	0.72	0.73	0.74
OQ total score	247	0.92	726	0.91	1309	0.93	2035	0.96	0.82	0.79

*62 cases.
OQ = Outcome Questionnaire.

We tried replacing missing values with the mean score of the remaining scale items. This resulted in slightly better *a* values for the clinical sample, but not for the community and the university samples. Beside restriction of range, the SR subscale has the lowest number of items, and two of the items that were awkward in the three-factor factor analysis belong to this scale. These two items, items 14 and 32, have relatively low item-total correlations (r_{it} = 0.11–0.15).

Most values are similar to values that were found in the American sample. No reports of internal consistency in the American community sample exist, so comparison is not possible, but in a German community sample, Cronbach's alpha for the SR subscale has been found to be 0.59, which is close to our value of 0.53 (Lambert, Hannöver, Nisslmüller, Richard, & Kordy, 2002).

Test–retest reliability is an indication for the stability of scoring over time. Very marked score changes over a short period of time would be problematic. The correlation between the first and second completion of the OQ is sufficient for both clinical (r_{tt} = 0.70–0.83) and student (r_{tt} = 0.71–0.81) samples.

Validity

Criterion Validity. An important validity requirement of an outcome measure is that it should discriminate between the clinical population for which it is designed and the functional (community) population. Table 8 shows that the difference between community and clinical means were large. The community sample has a highly significantly better level of functioning on all subscales and the total scale, Wilks' $\lambda = 0.58$, $F(5, 2637) = 388.1$, $p < 0.001$. Effect sizes for the difference between the clinical and community samples are very large for the IR ($F[1, 2643] = 960.5$, $p < 0.001$, $d = 1.32$) and SR ($F[1, 2643] = 674.1$, $p < 0.001$, $d = 1.10$) subscales, and huge for the SD ($F[1, 2643] = 1873.2$, $p < 0.001$, $d = 1.83$) and ASD ($F[1, 2643] = 1280.7$, $p < 0.001$, $d = 1.52$) subscales and total scale ($F[1, 2643] = 1804.5$, $p < 0.001$, $d = 1.80$).

Concurrent Validity. To assess the concurrent validity, three subsamples completed additional questionnaires together with the OQ. Results are presented in Table 8.

The SCL-90 and DASS were used to validate the SD and ASD subscales. The concurrent validity of the SD subscale with the GSI of the SCL-90 was slightly below the American value in the clinical sample ($r = 0.80$ versus $r = 0.84$), but better in the university sample (0.78 versus 0.61). The correlations between the SD and DASS subscales were adequate: neither too high, nor too low. The ASD subscale also showed proper concurrent validity with the SCL-90 and the Anxiety subscale of the DASS ($r = 0.74$). Correlations between the Depression ($r = 0.63$) and Stress ($r = 0.60$) subscales and the ASD subscale were lower, as was to be expected.

It was difficult to find a Dutch instrument to validate the IR and SR subscales, and we had to calculate our own indices with the instrument we finally used. Nonetheless, the convergent validity of the GVSG-45 with the IR ($r = 0.51$) and SR sub-

Table 8. Current validity estimates for the OQ with Symptom Checklist-90, DASS and GVSG-45

	Clinical (n = 118)				University (n = 268)		
	GSI	DASS-D	DASS-A	DASS-S*	GSI	FIR	FSR
Symptom Distress	0.80	0.78	0.74	0.72	0.78	0.42	0.54
Anxiety and Somatic Distress	0.75	0.63	0.74	0.60	0.66	0.34	0.42
Interpersonal Relations	0.62	0.54	0.38	0.54	0.59	0.51	0.51
Social Role	0.51	0.51	0.46	0.48	0.57	0.38	0.55
OQ total score	0.80	0.77	0.68	0.72	0.77	0.49	0.60

All correlations are significant at the 0.01 level.
*117 cases.
GSI = Global Severity Index of the Symptom Checklist-90-Revised. DASS-D = Depression Anxiety Stress-Depression subscale. DASS-A = Depression Anxiety Stress-Anxiety subscale. DASS-S = Depression Anxiety Stress-Stress subscale. FIR = functioning on interpersonal relationship, based on the GVSG-45-item version subscales Parents, Partner, Children and Friends. FSR = functioning on social role, based on the GVSG-45 (Groningen Questionnaire of Social Behaviour) subscales Study, Work, Housework and Leisure. OQ = Outcome Questionnaire. GVSG-45 = Groningse Vragenlijst Sociaal Gedrag-45. DASS = Depression Anxiety and Stress Scale.

Table 9. Correlations for the Outcome Questionnaire (OQ) with instruments measuring other psychological constructs

	n	Symptom Distress	Anxiety and Somatic Distress	Interpersonal Relations	Social Role	OQ total score
ACQ	119	0.58	0.62	0.27	0.38	0.56
BSQ	119	0.50	0.57	0.22	0.24	0.46
ICG-r	56	0.60	0.62	0.33	0.38	0.60
IESR	110	0.44	0.47	0.11*	0.13*	0.35
LSAS-SR	54	0.62	0.50	0.48	0.34	0.63
PI-R	137	0.57	0.52	0.31	0.35	0.55
PSWQ	122	0.38	0.26	0.15*	0.18*	0.33
QIDS-SR16	164	0.78	0.65	0.47	0.44	0.77

*Non-significant correlations
ACQ = Agoraphobic Cognitions Questionnaire. BSQ = Body Sensations Questionnaire. ICG-r = Inventory of Complicated Grief-revised. IESR = Impact of Events Scale Revised. LSAS-SR = Liebowitz Social Anxiety Scale-Self-Report. PI-R = Padua Inventory-Revised. PSWQ = Penn State Worry Questionnaire. QIDS-SR16 = Quick Inventory of Depressive Symptoms-Self-Report 16-item version.

scale (r = 0.55) falls in the range of correlations of the OQ subscales with the Inventory of Interpersonal Problems (r = 0.49–0.64) and SAS, respectively, in the American samples (r = 0.44–0.73).

Correlations with other Psychological Constructs. In the internet screening tool sample, several instruments that measure specific disorders were administered together with the OQ (see Table 9). On almost all specific questionnaires, validity estimates are good, showing high correlations with the SD subscale (and subsequently, the OQ total scale) and lower correlations with the IR and SR subscales. Exceptions are the PSWQ and the Quick Inventory of Depressive Symptoms-Self-Report 16 (QIDS-SR16). The correlation between the PSWQ and the SD subscale is not high (r = 0.38), even though it is lower for the ASD, IR and SR subscales, as expected. The concept of worrying may not be uniquely linked to a certain pattern in symptoms. The QIDS-SR16 shows a good concurrent validity with the SD subscale (r = 0.78), but correlations with the ASD (r = 0.65), IR (r = 0.47) and SR subscales (r = 0.44) seem higher than desirable. The ASD subscale shows good concurrent validity on the ACQ (r = 0.62) and BSQ (r = 0.57) and the LSAS-SR (r = 0.50). Unexpected is the relatively high correlation of the ASD subscale with the ICG-r (r = 0.62). Of further interest is the correlation between the IR subscale and the LSAS-SR (r = 0.48). Having symptoms of social anxiety will probably influence interpersonal functioning, which shows a subsequently somewhat higher correlation.

In summary, the concurrent validity of the SD, ASD subscales and the total scale seems good, but validity estimates of the IR and especially SR subscales are less convincing (Table 9).

Sensitivity to Change. Another important criterion for instruments that are used for outcome and

DOI: 10.1002/cpp

progress research is that they are capable of measuring changes in functioning that occur as a result of treatment. A subsample of 60 patients received a short treatment, with a maximum of five sessions. The OQ was administered before and after treatment. The OQ showed high sensitivity to change on all subscales (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: $t(55) = 4.1$, $p < 0.001$, $d = 0.77$) and the total scale, $t(56) = 7.1$, $p < 0.001$, $d = 1.33$.

DISCUSSION

This study investigated the cross-cultural validity of the OQ in the Dutch population by evaluating the psychometric properties of the Dutch version and its equivalence with the original version of the OQ. The results show that the language versions are similar when it comes to reliability and validity estimates, but differences in factor structure and normative scores have been found.

The three-domain structure of the instrument, for which there was no strong evidence in the original version, had a reasonable fit in the Dutch population. Further analyses resulted in two additional factors that overlap mainly with the SR and SD subscales. The first one, which consisted of four items that are in the SR domain, was unexpected but not unexplainable considering the bad performance of item 14 ('I work/study too much'). This item has low item-total correlation and also came to notice in the reliability analysis. It is problematic in the original OQ as well (see Mueller et al., 1998) and probably does not represent problematic behaviour. In fact, in contemporary society, some people may consider working too hard a good quality.

The second factor, named ASD, was considered a useful addition to the existing scales and was therefore used in further analyses. Reliability and validity estimates for the ASD factor are promising. This factor may be especially interesting for use by care providers that specialize in anxiety or psychosomatic disorders.

Finding additional factors to the original structure does not seem to indicate conceptual equivalence between the two versions of the OQ. And the fit of our solution is notably better than the fit of the original three-factor solution. However, this does not necessarily imply that they are not equivalent. Some of the GFIs that we found for the three-factor solution were similar to the ones reported in the American sample (Mueller et al., 1998). Running the same statistical analyses as we did may result in a similar structure in the American OQ.

The correlations between the subscales were too high. This suggests inadequate conceptual equivalence. Another possible explanation for the high correlations between the subscales may be that there really is a mutual interdependence between the concepts and that distress in one area influences functioning in other areas.

Comparison of normative scores between the American and Dutch populations showed that the Dutch community and clinical samples scored somewhat below their American equivalents. As was mentioned earlier, differences in scoring between culturally different populations are common in psychological testing. For instance, differences were found in the EPSILON study and on instruments such as the MMPI. The Dutch students showed higher scores than the American students. This may be due to a difference in sampling. In the Netherlands, the sample consisted of psychology students, whereas the American sample included other disciplines as well. Even though the differences between the populations were relatively small, calibration of cut-off scores and RCIs was necessary, for a lack of criterion equivalence occurred. Calibration resulted in a cut-off score for the Dutch population of eight points below the American cut-off point. After calibration, sensitivity and specificity values were very similar to those of the original version. The specificity of the American and Dutch OQ is 0.83 and 0.85, respectively, and the sensitivity is 0.84 for both versions. The RCIs were equal as well.

A marked difference between the American and Dutch normative scores is that in the Dutch population, gender differences were found in both the clinical and the community sample. Men had more problems in the SR domain, whereas women showed higher levels of SD as well as ASD. Gender differences in normative scores are quite common in testing, and it is surprising that they were not found in the original OQ. The OQ manual reports that in one study, some statistically different mean scores were found in the patient sample, but they were not considered to be of clinical relevance. In the OQ manual, gender scores are only reported on a relatively small subsample, so a lack of power may be causing the insignificant findings.

When developing normative scores for any test, the quality of the population samples is very important. By combining phone book and business setting sampling, we strived for a representative sample of the Dutch functional population. This

DOI: 10.1002/cpp

was more complicated for the clinical sample, as we could only address patients who received treatment in the participating mental health care centres. Given that the sample size is large and multiple mental health care organizations participated, we believe our sample to be representative for the Dutch outpatient population.

Besides examining the equivalence, the psychometric properties of the Dutch version were investigated. The reliability of the subscales and the total scale was adequate in most of the samples. An exception was the internal consistency of the SR domain, which was too low in all three samples, but was substantially better when the clinical and community samples were combined. Sensitivity to change is very good, and the OQ can effectively discriminate between functional and dysfunctional populations. The concurrent validity showed proper values for the SD and ASD subscales, but less support for the IR and SR subscales.

This study did not address the OQ as a measure for tracking patient progress. More research should be conducted with the Dutch OQ on this subject to obtain a better comparison of progress curves between the Dutch and American population. Given the differences in normative scores that were found in the current study, differences in treatment progress are to be expected. In connection with that, the sensitivity to change on a session to session basis should be investigated. Also, further research should be conducted on the SR domain and the additional factor that was identified in the present study. We are currently performing a pilot study with a different formulation of item 14 ('My work/study is too much for me') that may better reflect problematic functioning in the SR domain. The Dutch OQ has moderate to good psychometric properties and is ready for use in clinical practice. However, separate norms for patient progress should be developed for the Dutch population.

Summarized, the Dutch OQ has similar psychometric properties as the original instrument, but the two versions are not equivalent on all aspects. There may be a difference in conceptual equivalence, although further analyses with the original instrument can prove otherwise. Criterion validity of the Dutch OQ was similar to the original values, but only after calibration of cut-off scores, indicating a lack of criterion equivalence. These results imply that a similarity in psychometric properties does not guarantee equivalence. The fact that calibration of cut-off scores was necessary, even though differences in population scores were small, shows the importance of proper normative scores for translated instruments. This is especially true for clinical outcome measures such as the OQ, where use of the 'wrong' norms may lead to faulty treatment decisions.

REFERENCES

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

Boelen, P.A., de Keijser, J., & van den Bout, J. (2001). Psychometrische eigenschappen van de Rouw Vragenlijst (RVL). *Gedrag & Gezondheid*, *29*(2), 172–185.

Bouman, T.K. (1995). De Agoraphobic Cognitions Questionnaire (ACQ). *Gedragstherapie*, *27*, 69–72.

Bouman, T.K. (1998). De Body Sensation Questionnaire (BSQ). *Gedragstherapie*, *31*, 162–168.

Brom, D., & Kleber, R.J. (1985). De schokverwerkingslijst. *Nederlandsch Tijdschrift voor Psychologie*, *40*(3), 164–168.

Butcher, J., Derksen, J., Sloore, H., & Sirigatti, S. (2003). Objective personality assessment of people in diverse cultures: European adaptions of the MMPI-2. *Behaviour Research and Therapy*, *41*, 819–840.

Chambless, D.L., Caputo, G.C., Bright, P., & Gallager, R. (1984). Assessment of fear of fear in agoraphobics: The Body Sensations Questionnaire and the Agoraphobic Cognitions Questionnaire. *Journal of Consulting and Clinical Psychology*, *52*, 1090–1097.

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

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.

de Beurs, E., van Dyck, R., Marquenie, L.A., Lange, A., & Blonk, R.W.B. (2001). De DASS: Een vragenlijst voor het meten van depressie, angst en stress. *Gedragstherapie*, *34*(1), 35–53.

de Jong, A., & van der Lubbe, P.M. (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.

de Jong, K., & Nugter, M.A. (2004). De Outcome Questionnaire: Psychometrische kenmerken van de Nederlandse vertaling. *Nederlandsch Tijdschrift voor Psychologie*, *59*(3), 76–79.

Derogatis, L.R. (1975). *The brief symptom inventory*. Baltimore, MD: Clinical Psychometrics Unit.

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

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clininal outcomes in routine evaluation. *Journal of Mental Health*, *9*, 247–255.

DOI: 10.1002/cpp

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*, 257–263.

Gladis, M., Gosch, E.A., Dishuk, N.M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*(3), 320–331.

Hatfield, D.R., & Ogles, B.M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology—Research and Practice*, *35*(5), 485–491.

Hofstede, G. (2006). Cultural dimensions. Retrieved March, 2006, from http://www.geert-hofstede.com/hofstede_dimensions.php

Interapy. (2004). Interapy Nederland B.V. Amsterdam: The Netherlands.

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*(1), 12–19.

Jacobson, N.S., Follette, W.C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*(4), 336–352.

Jacobson, N.S., Roberts, L.J., Berns, S.B., & McGlinchey, J.B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300–307.

Kline, R.B. (1998). *Principles and practice of structural equation modeling*. New York: 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 and Psychotherapy*, *3*(4), 249–258.

Lambert, M.J., Whipple, J.L., Smart, D.W., Vermeersch, D.A., Nielsen, S.L., & Hawkins, E.J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research*, *11*(1), 49–68.

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*(1), 40–46.

Lambert, M.J., Whipple, J.L., Hawkins, E.J., Vermeersch, D.A., Nielsen, S.L., & Smart, D.W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, *10*(3), 288–301.

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)* (3rd ed.). Wilmington, DE: American Professional Credentialling Services LLC.

Lambert, M.J., Harmon, C., Slade, K., Whipple, J.L., & Hawkins, E.J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. *Journal of Clinical Psychology*, *61*(2), 165–174.

Liebowitz, M.R. (1987). Social phobia. *Modern Problems in Pharmacopsychiatry*, *22*, 141–173.

Lovibond, S.H., & Lovibond, P.F. (1995). *Manual for the Depression Anxiety Stress Scales*. Sydney: Psychology Foundation of Australia.

Meyer, T.J., Miller, M.L., Metzger, R.L., & Borkovec, T.D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *28*, 487–495.

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.

Prigerson, H.G., Kasl, S.V., & Jacobs, S.G. (1997). *The Inventory of Complicated Grief Revised*: Unpublished manuscript.

Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B.A., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., & Keller, M.B. (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS) Clinician Rating (QUIDS-C) and Self-Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*, 573–583.

Sanavio, E. (1988). Obsessions and compulsions: The Padua Inventory. *Behaviour research and Therapy*, *26*, 167–177.

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.htm

van Balkom, A.J.L.M., de Beurs, E., Hovens, J.E.J.M., & van Vliet, I.M. (2004). Meetinstrumenten bij angststoornissen. *Tijdschrift voor psychiatrie*, *10*, 687–692.

van der Ploeg, E., Mooren, T.T.M., Kleber, R.J., van der Velden, P.G., & Brom, D. (2004). Construct validation of the Dutch versions of the Impact of Events Scale. *Psychological Assessment*, *16*(1), 16–26.

Van Oppen, P., Hoekstra, R.J., & Emmelkamp, P.M.G. (1995). The structure of obsessive compulsive symptoms. *Behaviour Research and Therapy*, *33*, 15–23.

van Rijsoort, S., Vervaeke, G., & Emmelkamp, P. (1999). The Penn State Worry Questionnaire and the Worry Domains Questionnaire: Structure, reliability and validity. *Clinical Psychology and Psychotherapy*, *6*, 297–307.

van Wijngaarden, B., Schene, A.H., Koeter, M., Vazquez-Barquero, J.-L., Knudsen, H.-C., Lasalvia, A., McCrone, P., & The EPSILON Study Group. (2000). Caregiving in schizophrenia: Development, internal consistency and reliability of the Involvement Evaluation Questionnaire–European Version: EPSILON Study 4. *British Journal of Psychiatry*, *177*(Suppl. 39), S21–S27.

Weiss, D., & Marmar, C. (1997). The impact of event scale—revised. In J. Wilson & T. Keane (Eds), *Assessing psychological trauma and PTSD* (pp. 399–411). New York: Guilford.

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

DOI: 10.1002/cpp