

# Chapter 3

## Validation of Two Outcome Measures for Routine Outcome Monitoring in Psychiatry: the HORVAN Study

V.J.A. Buwalda, S. Draisma, J.H. Smit, J.A. Swinkels, W. van Tilburg  
*Tijdschrift voor Psychiatrie 2011; 53: 715-726*

## Chapter 3 | Validation of Two Outcome Measures

### Abstract

*Background:* Transparency in psychiatry can be increased by the use of routine outcome monitoring (ROM) instruments. Instruments should be easy to use and take very little time to complete; they also need to have psychometric qualities, **be sensitive to change, and provide information about patients' symptoms, and** about interpersonal and social functioning.

*Aim:* To investigate to what extent the Dutch translation of the combination of the Health of the Nation Outcome Scales (HoNOS) and Outcome Questionnaire (OQ) meets the above-mentioned quality criteria and to examine how the combination relates to the Symptoms CheckList (SCL-90).

*Method:* Data for 148 patients collected at three measurement time-points were available for analysis. The psychometric qualities of the instruments and their sensitivity to change were carefully monitored.

*Results:* The three scales showed high values for internal consistency (Cronbach's alpha). The HoNOS total score and the subscales of the OQ correlated reasonably well with the SCL-90 total score (convergent validity). At the first measurements, patients with a comorbid diagnosis had the lowest scores (discriminant validity). The clinically significant change between T1 and T2 and between T2 and T3 was sufficiently high for all measurement instruments.

*Conclusion:* The combination of the HoNOS rating scale and the self-report list OQ seems to be suitable for ROM in psychiatry.

**Keywords:** clinically significant change (CSC) – routine outcome monitoring (ROM) – validation study

## 1. Introduction

In recent decades, there has been an increasing demand for transparency in the clinical-care process in the Dutch healthcare system. Outcome measures have been developed that clarify the outcomes of individual treatment or measure patient satisfaction. In mental healthcare, Routine Outcome Monitoring (ROM) is used (De Beurs & Zitman, 2007). This method involves the collection of data on the functioning and well-being of the patient before, during, and after treatment using self-rating scales and/or observer-rating scales. Treatment can be adjusted when improvement proves to be insufficient (Aartsen et al., 2010).

There are various useful outcome measures to evaluate the course of treatment, but only a few meet the requirement of *validly* measuring improvement (Hermann, 2005; Lambert et al., 2001; Thornicroft et al., 2005).

The authors above have formulated the following quality criteria for ROM measures:

- (1) They are quick and easy to use;
- (2) They have a minimum of three distinct domains: symptoms, interpersonal problems, and social-role performance (Lambert et al., 2001);
- (3) They have sufficient psychometric quality; and
- (4) They are sensitive to changes in the short term.

## SCL-90

The most widely applied and studied outcome measure used to assess treatment effects in clinical practice is the Symptoms CheckList (SCL-90), which is a self-rating scale that embraces the entire spectrum of disorders (Arrindell & Ettema, 1986; Arrindell & Ettema, 2003; Derogatis, 1977). However, the SCL-90 has a number of disadvantages:

- 1) a large group of seriously ill patients suffering from concentration deficiencies, cognitive disorders, or language problems cannot be included because they are not able to complete the questionnaire themselves;
- 2) items regarding social interaction and social-role performance are lacking; and
- 3) the questionnaire takes a relatively long time to complete.

## Chapter 3 | Validation of Two Outcome Measures

### HONOS and OQ

Outcome measures that do not have these shortcomings are the *Health of the Nation Outcome Scales* (HoNOS) and the *Outcome Questionnaire* (OQ). These are, respectively, an observer-rating scale aimed at more severe psychiatric disorders, psychotic disorders, and bipolar disorders (Sharma et al., 1999) and a self-rating scale aimed at symptoms of anxiety and depression (Lambert et al., 2001). In the **HoNOS, OQ, ROM Validity [The] Netherlands** Study (the HORVAN Study), we investigate the HoNOS and OQ in a ROM-configuration in a Dutch research population of outpatients. In addition, we verified the user-friendliness, psychometric quality, and sensitivity to change. This ROM configuration is compared to the SCL-90 because the latter is a widely used instrument and its use is generally recommended as the “gold standard” (Oiesvold, 2010).

The two questions this study aims to answer is: *To what extent does the combination of HoNOS and OQ serve the Dutch situation with respect to the four requirements for the quality of ROM outcome measures and how does the combination HoNOS-OQ relates to the SCL-90?*

First we will discuss the characteristics of the three outcome measures and then we will outline the method and results of the HORVAN Study.

## 2. The Outcome Measures and Quality Criteria

### 2.1 User-friendliness

The HoNOS is an observer-rating scale with 12 items completed by the clinician **concerning the patient and based on the patient's case history over the previous two weeks**. The HoNOS was developed in England by Wing et al. (1998), and it is regularly used internationally (Slade et al., 1999; Trauer et al., 1999). It takes very little time to complete, and the instrument can be used without disorder-specific restrictions. The clinician will, of course, require training in the assessment procedure. The OQ is a self-rating scale containing 45 items for the domains “symptoms,” “interpersonal functioning,” and “social role performance” during the previous week. The OQ was developed in the United States (Lambert, 1996). The *Symptom Checklist-90* (SCL-90) is a much-used, multi-dimensional self-report measure developed in the US by Derogatis (1977). With its 90 items, the list endeavors to represent the complete variety of possible psychiatric symptoms, and therefore the time to complete is longer than the HoNOS or the OQ.

The different sub-scales of the three outcome measures are shown in Table 1.

## Validation of Two Outcome Measures | Chapter 3

**Table 1.** Sub-scales of three outcome measures (number of items) that were used in the HORVAN study (HoNOS, OQ, and SCL-90)

Name of the scale and number of items	
Sub-scales	<b>Health of the Nation Outcome Scales (HoNOS) (12 items)</b>
	1. Cognitive and physical impairment (2)
	2. Behavior (3)
	3. Symptoms (3)
	4. Social problems/relationships/daily activities (4)
	<b>Outcome Questionnaire (OQ) (45 items)</b>
	1. Interpersonal relations (11)
	2. Social role performance (9)
	3. Symptoms (25)
	<b>Symptom CheckList (SCL-90) (90 items)</b>
	1. Anxiety (10)
	2. Agoraphobia (7)
	3. Depression (16)
4. Somatization (12)	
5. Inadequate thought and action (9)	
6. Mistrust and interpersonal sensitivity (18)	
7. Hostility (6)	
8. Sleeping disorders (3)	
9. Psychoticism (9)	

Table 2 illustrates the use and other criteria for these outcome measures. It is clear that the OQ best meets the requirement of including the three domains (symptom distress, interpersonal relations, and social role performance). While the HoNOS also includes four items on daily and social functioning in addition to psychiatric symptoms, the SCL-90 **emphasizes the severity of the respondent's** symptoms.

**Table 2.** Characteristics of three ROM measures

	<b>HoNOS</b>	<b>OQ</b>	<b>SCL-90</b>
<b>Disorder</b>	Psychosis & bipolar disorder	Anxiety & mood disorders	All psychiatric disorders
<b>Number of items</b>	12 (possibly 3 additional items)	45	90
<b>Likert scale</b>	0-4	0-4	1-5
<b>Likert scale</b>	(no-problem-severe to extremely severe problem)	(never – nearly always)	(not at all – very serious)
<b>Type of measure</b>	Professional assessment	Self assessment	Self assessment
<b>Number of subscales</b>	4	3	9
<b>Time required (minutes)</b>	5	10	15-20
<b>Training required</b>	Yes	No	No
<b>Availability</b>	Free to download	Payment required	Payment required
<b>Reliability</b>	0.78 (Mulder et al., 2004)	0.94 (Chapman, 2003)	0.73 – 0.97, 0.73 – 0.97 various samples (Arnindell and Ettema, 2003)
<b>Cronbach's alpha</b>			
<b>Inter-observer reliability</b>	0.29-0.92 per item (Ornel et al., 1999) 0.71-0.96 per various samples (Trauer et al., 1999)	0.79 (Chapman 2003)	-
<b>Convergent validity (R other outcome measures)</b>	0.92 (Mulder et al., 2004) 0.92 (Broersma & Sytema, 2008) Reasonable with GAF, QCL, MANISA, CANSAS, CGI, BPRS (Mulder et al., 2004) Good with BPRS and RFS (Wing et al., 1998)	Reasonable to good with (BDI, STAI, SF36, SCL90) (Umphress et al., 1997) Reasonable with sub-scales-BASIS-32 (Doefler et al., 2002) Beurs et al., 2004 Jong et al., 2007 Total score < 35	High with GHQ – 12 (Holl et al., 2003) High with GHQ (Koeter, 1992)
<b>Control/Community groups</b>	-		Derogatis, 1997 Arnindell & Ettema, 1986
<b>CSC cutoff</b>	All items with values < 2	Total score < 35	Senscore Total score < 124

## Validation of Two Outcome Measures | Chapter 3

### *2.2 Psychometric quality*

In The Netherlands, the HoNOS is considered to be reliable and valid by Aartsen et al. (2010), Broersma and Sytema (2008), and Mulder et al. (2004). For the construct validity, in terms of distinctiveness, Mulder et al. compared scores between clinical departments. Admitted patients scored significantly higher than outpatients; patients in day care scored in between these two groups. Broersma and Sytema replicated these results for geriatric psychiatry. In addition, patient groups with psychotic, depressive, and anxiety disorders were compared with one another as well: individual items and subscales showed significant differences between these groups. Mulder et al. found reasonable convergent validity in the relationship between the HoNOS subscales and the total scores of the other outcome measures (*Global Assessment of Functioning* (GAF), *Manchester Short Assessment of Quality of Life* (MANSA), *Camberwell Assessment of Need Short Appraisal Schedule* (CANSAS), *Clinical Global Impression Scale* (CGI), and the *Brief Psychiatric Rating Scale* (BPRS)).

There is sufficient research available on the psychometric quality of the OQ (e.g., De Beurs et al., 2005; De Jong et al., 2007; Doerfler, 2002; Umphress et al., 1997). The reliability of both the American and the Dutch OQ is good (Chapman, 2003; De Beurs et al., 2005; De Jong et al., 2007).

Umphress et al. (1997) demonstrated with different OQ total scores that a group of psychiatric patients could be easily distinguished from a group within the general healthy population. De Jong et al. (2007) replicated this study for The Netherlands. Umphress et al. also found good convergence with the *Beck Depression Inventory* (BDI), the *Spielberger State-Trait Anxiety Inventory* (STAI), the *Short Form* (SF-36), and the SCL-90.

Doerfler et al. (2002) found reasonable divergent and convergent validity of the OQ sub-scales with those of the *Behavior and Symptom Identification Scale* (BASIS-32) subscales (psychiatric symptoms and social functioning). The subscales differentiated clearly between various diagnostic groups.

Arrindell and Ettema (1986) used various samples to measure the internal consistency of the SCL-90 scales. **Cronbach's alphas for the different subscales** varied from 0.73 to 0.97. Holi et al. (2003) found significantly different average scores between samples from the general healthy population and outpatients. Koeter (1992) compared the depression and anxiety scales of the SCL-90 with those of the *General Health Questionnaire* (GHQ-28) and compared these to a

## Chapter 3 | Validation of Two Outcome Measures

DSM-III diagnosis in a sensitivity and specificity analysis. The concurrent validity of the subscales of the SCL and the GHQ were found to be good.

### *2.3 Sensitivity to change*

For measuring change, clinically significant change (CSC), the method of Jacobson and Truax, is available. This implies that (1) a statistically reliable change (RC) occurs and (2) if a cut-off score is exceeded, which marks the transition from illness to recovery (CSC), the score of the (dysfunctional) patient should exceed the standard score of a functional normal population. If only the first criterion is met, then there is **reliable improvement, “but not yet” recovery**. If only the second criterion is met, then there is a shift from dysfunctional or ill to functional (healthy), but both scores are so close to the cut-off score that the shift has no clinical significance (Aarsse et al., 2003; De Beurs et al., 2005).

Parabiaghi et al. (2005) studied the CSC of the HoNOS in an Italian population. The authors, however, used an adapted algorithm without a specific cut-off score for distinguishing between clinical and functional. As long as a patient does not score a 2 or higher on any single item, he could be considered as subclinical (functional). In the sample, 5.6% recovered, and in a subgroup of patients with severe problems, 14.4% recovered. Other authors measured changes over time with the conventional statistical analysis techniques. Wing et al. (1998) tested the differences between the HoNOS scores with t-tests on two separate occasions. Hunter et al. (2009) used regression analyses for patients with schizophrenia and found little change. All the studies that we reviewed on change with the OQ are based on t-tests and were not worked out according to the CSC method. Doerfler et al. (2002) and Lambert (1996) found significant improvements on all subscales between patient admission and discharge. Vermeersch et al. (2000) studied sensitivity to change at the item level.

Arrindell and Ettema (2003) show an overview of SCL-90 studies from 1957 to 1998 of the effects of short as well as long-term treatment. In all the studies discussed, change is determined by conventional statistical techniques, not with CSC. Koeter (1992) used results of a sample of outpatients on three measurement points but could not identify explicit change. Wilson et al. (1997) did use the CSC to determine the recovery of patients with PTSD treated with eye movement desensitization and reprocessing (EMDR) therapy by using the SCL-90: after 15 months, 56% showed a clinically significant improvement.

We can conclude that there has been little research into improvement assessed by the CSC method with respect to the three measures.

### *3. HORVAN Study: Method*

#### *3.1 Procedure and sample*

The population studied consisted of patients from a medium-sized Dutch town **who had been referred to an outpatients' clinic by their primary care practitioners.** These patients were interviewed initially by a resident in psychiatry or a psychologist upon intake. Subsequently, a structured interview of 30-40 minutes was carried out by a supervised resident in psychiatry or a psychiatrist using the MINI International Neuropsychiatric Interview (MINI; Van Vliet & De Beurs, 2007; Sheehan et al., 1998). The MINI is used to diagnose DSM-IV psychiatric disorders Axis-I diagnosis. In addition to the HoNOS, OQ and SCL-90 were assessed.

**One of the authors (V.J.A.B.) works as a psychiatrist at the outpatients' clinic and,** after training, assessed the patients using the HoNOS from January 2002 to November 2004. In addition, the author also trained a resident in psychiatry, working under his supervision, to score the HoNOS. The HoNOS scores were based on interviews with the patients after the assessment of the MINI, supplemented with information given by the patients and on information from **the patients' clinician.** Subsequently, **the patients completed the OQ and the SCL-90.** During treatment, on average every 10 weeks, the patient was asked to complete the OQ and the SCL-90; the HoNOS, GAF, and CGI were completed by the resident in psychiatry or the psychiatrist. In this way, changes in the severity of symptoms were monitored during treatment.

In total, 370 patients were seen for a baseline assessment in which at least three outcome measures were completed. Some of the patients took part in follow-up measurements (N=213, 54%) during a treatment period of (at most) 2 years and 6 months, with a maximum of 17 measurements from a small group of patients. The aim was to assess patients every six weeks in cases where medication was used as the treatment. Where cognitive behavioral therapy was the chosen treatment, the assessment took place once every three months. A small group of patients (N=11.3%) was admitted to the clinic during treatment. The drop-out reasons for 157 people who did not show up for the follow-up measurement (T2) were: non-appearance at appointments, receiving follow-up treatment outside **the outpatients' clinic, an incomplete pre-test measurement, no diagnosis, or being placed on a waiting list for treatment.**

## Chapter 3 | Validation of Two Outcome Measures

Finally, data from 148 patients were found to be useful for statistical analysis for the first three outcome measures. Of these patients, at least one follow-up measurement was available at time point 2 (T2) with the three outcome measures. At the third measurement time point, the results of at least one of the three outcome measures was missing from 28% of these individuals (N=41). Missing data for these individuals on the third measurement were supplemented **by imputation according to the “intention to treat, last value carried forward”** principle.

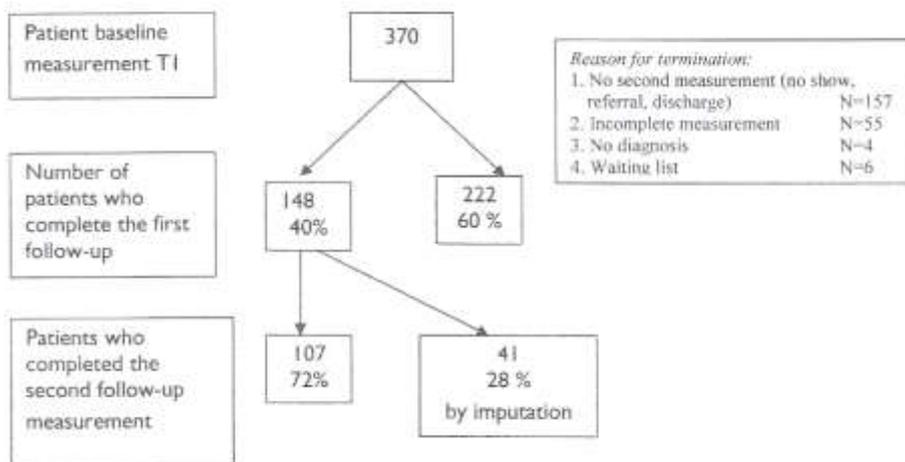


Figure 1. Flowchart of participants

### 3.2 Analysis

Cronbach's alphas were computed for reliability; correlations between scales for convergent validity, t-tests, CHI-square tests, and ANOVA analyses of variance were used to indicate differences between groups. Reliable change was calculated using the formula  $RC = (X_{t1} - X_{t0}) / S_{diff}$  [where  $S_{diff} = \sqrt{2 (SE)^2}$ ], i.e., changes expressed in the difference score were corrected for the standard error. To calculate the standard error, standard deviations and Cronbach's alphas from the first measurements were used. Subsequently, we examined how many individuals with an RC also exceeded the norm score from dysfunctional to functional (healthy).

In order to calculate CSC, norms for the OQ were available, with a score lower than 55 indicating functional (healthy) (De Jong et al., 2007, p. 296); for the HoNOS and the SCL-90, these scores were indirectly available. For the HoNOS, this means that there is no item with a value greater than 1 (Parabiaghi et al., 2005);

for the SCL-90 this means a total score <124 (Arrindell & Ettema, 1986, p. 39, average general population). To determine the suitability of a combined OQ-HoNOS, a sensitivity analysis was carried out using the SCL-90 as the gold standard.

**4. Results**

*4.1 Research sample: completers versus non-completers*

Of 148 individuals, data for the first two measurements were complete. By the third measurement, data for 41 of the 148 individuals was missing. For these 41, **data on the T3 were imputed (according to the “last value carried forward” principle)**. Multiple imputation was also applied, with demographic variables and scores of the first two measurements as predictors. This yielded no significant differences. The most important demographic characteristics of the sample are presented in Table 3. Clinical diagnoses were developed by re-grouping the DSM-IV diagnoses into four main categories.

**Table 3.** Socio-demographic characteristics of the HORVAN sample

Characteristics	Abs.	%
Gender		
Male	53	36 %
Female	92*	62 %
Education		
Low	87	74 %
High	30**	26 %
Primary clinical diagnosis		
Anxiety	31	21 %
Depression	34	23 %
Comorbid anxiety and depression	59	40 %
Other diagnoses: psychosis (14), addiction (5), uncertain (5)	24	16 %
	mean	St.dev. (range)
Age	37.7	13.1 (16-67)

\* 3 individuals' gender is unknown;

\*\* 31 individuals' education level is unknown. Higher level of education cut-off is from eleven years of education

## Chapter 3 | Validation of Two Outcome Measures

The average age was 37.7; the majority of patients were female and highly educated. The completers (N=148, at least 2 measurements) were compared on available variables with the non-completers (N=222, only baseline measurements). There were no differences between completers and non-completers (follow-up) in age ( $T_{11} t_2 = 0.30$ ,  $df=256$ ,  $P=0.77$ ), gender distribution ( $\chi^2_{t1_t2}=1.17$ ,  $df=1$ ,  $p=0.32$ ), or distribution over the main diagnostic groups ( $F_{t1_t2}=5.7$ ,  $df=1$ ,  $p=0.02$ ). In addition, average (sub-scale) scores on three outcome measures did not vary significantly between completers and non-completers.

### 4.2. Reliability

In Table 4, Cronbach's alphas are shown for the OQ, HoNOS and SCL-90 at three measurement time points.

**Table 4.** Reliability coefficients (Cronbach's alpha) of the outcome measures

Subscales (n items)	$\alpha^*$	$\alpha$	$\alpha$
	T1**	T2	T3
OQ Symptom distress (25)	0.90	0.93	0.96
OQ Social role (8)	0.60	0.69	0.78
OQ Interpersonal relations (11)	0.76	0.82	0.89
OQ Total (45)	0.92	0.95	0.97
HoNOS Behavior (3)	0.31	0.38	0.37
HoNOS Impairment (2)	0.55	0.53	0.32
HoNOS Symptoms (3)	0.24	0.32	0.42
HoNOS Social (4)	0.49	0.60	0.53
HoNOS Total (12)	0.64	0.74	0.73
SCL Total (90)	0.97	0.98	0.98

\* Split half reliability coefficients give comparable values and sequences.

\*\* Mean number of days T1-T2= 75 (SD=59) T2-T3=72 (SD=55).

Three of the four OQ scales attained high values for internal consistency (Cronbach's alpha >.70). Only "social role" had an alpha lower than .65. This subscale contained only 9 items and provided the lowest reliability values, as was also found by other authors. For the HoNOS total scale, we found reasonably comparable alphas, commensurate with the 0.78 found by Mulder et al. (2004). The subscales of the HoNOS did not reach any acceptable alpha score, but the number of items was so small that this would have been quite difficult to achieve. The HoNOS is not primarily based on subscales with correlating items; instead, each individual item represents a complete dimension. For the SCL-90, all

## Validation of Two Outcome Measures | Chapter 3

reliability coefficients were high (subscales were more than sufficient: 0.76 to 0.94, inclusive). This was consistent with other studies.

### *4.3 Convergent validity*

Emulating Umphress et al. (1997), we explored convergent validity by comparing the sub-scales of the OQ and the HoNOS with those of the SCL-90 total score. Correlations between the scales of the various measures at baseline are shown in Table 5.

**Table 5.** Subscale correlations OQ45, HoNOS, and SCL-90 total score, etc.

Sub-scales	SCL-90 T1	SCL-90 T2	SCL-90 T3
OQ45: Symptom distress (OQ SD)	0.84	0.91	0.82
OQ45: Social role (OQ SR)	0.51	0.63	0.61
OQ45: Interpersonal relations (OQ IR)	0.63	0.71	0.68
OQ Total	0.84	0.90	0.90
HoNOS: Total	0.64	0.75	0.80

All correlations are significant at  $p < 0.01$ .

Subscales of the OQ correlated sufficiently enough with the SCL-90 total score to conclude good convergent validity; the scales indicated a similar construct. High correlations of the OQ and the HoNOS with the SCL-90 supported the idea that both measures used together are interchangeable with the SCL-90.

## Chapter 3 | Validation of Two Outcome Measures

**Table 6.** Mean score per diagnosis and outcome measure

		OO*	HoNOS*	SCL-90*
		mean (range)		
Anxiety (N=31)	T1	68.8 (33-109)	5.6 (0-13)	177.7 (90-296)
	T2	58.6 (12-113)	4.4 (0-13)	165.2 (91-325)
	T3	55.3 (6-113)	4.1 (0-10)	156.2 (90-264)
Depression (N=34)	T1	69.4 (26-131)	6.4 (1-21)	185.3 (120-289)
	T2	68.1 (22-126)	5.6 (0-21)	176.2 (101-355)
	T3	56.7 (6-130)	4.8 (0-15)	156.8 (91-316)
Comorbid A-D (N=59)	T1	83.0 (49-131)	8.8 (1-21)	227.8 (137-347)
	T2	67.3 (1-124)	6.3 (0-20)	183.7 (91-317)
	T3	63.2 (12-137)	5.8 (0-19)	176.8 (90-384)
<i>Rest (N=24)</i>	T1	68.9 (24-140)	8.3 (0-23)	196.2 (102-344)
	T2	68.6 (25-119)	6.1 (2-13)	182.6 (106-330)
	T3	61.7 (24-124)	4.4 (1-16)	167.9 (114-347)

\* Only the average T1 scores differ significantly for diagnosis ( $F_{OO}=4.4$   $p=0.004$ ;  $F_{HoNOS}=8.2$   $p=0.01$ ;  $F_{SCL}=7.7$   $p=0.00$ ).

### 4.4 Discriminant validity

For the various diagnostic groups, differing averages and changes in values were expected. The prognosis for improvement in individuals with mood disorders is more positive than for those with a psychosis. To the extent that numbers in the diagnostic groups permitted, we investigated differences in averages for the three instruments (Table 6).

The average OO score was significantly higher for comorbid anxiety and depression than for other diagnoses on the first measurement. This higher score disappeared on measurements 2 and 3; it appeared that these benefitted highly from their treatment. This pattern could also be seen in the HoNOS; patients suffering from both anxiety and depression began with the most unfavorable score, and these scores decreased the most over the measurements. Even the **diagnosis for the category “other” began with a relatively unfavorable HoNOS score and dropped considerably.** This pattern was also seen in the results of the SCL-90.

## Validation of Two Outcome Measures | Chapter 3

### 4.5 Clinically Significant Change (CSC)

Table 7 illustrates percentages with reliable change (RC, at the relevant measurement points) and percentages that also clinically improved, surpassing the norm (CSC).

**Table 7.** Change between measurements T1-T2 and T2-T3 (RC = Reliable Change, CSC = clinical significant Change)

Measures	OQ45	HoNOS	SCL-90
RC T1-T2	23.6 %	7.4 %*	43.9 %
CSC T1-T2	18.9 %	3.4 %	12.8 %
RC T1-T3	37.2 %	11.5 %	54.1 %
CSC T1-T3	28.4 %	8.1 %	23.0 %

Expressed in terms of CSC to OQ scores, improvements were seen in 1 in 5 individuals between the first and second measurements and in 1 in 3 individuals between the first and third measurements. In short, the course of treatment measured with the OQ was more favorable over time and as treatment progressed. This was also found to a lesser extent for the HoNOS and the SCL-90.

### 4.6 SCL-90 as the gold standard

A sensitivity analysis was conducted on the OQ and the HoNOS using the SCL-90 as the gold standard. The norms mentioned in paragraph 4.5 were once again used as cut-off values. The data from the third measurement were used because these had the relatively largest number of healthy people (27.1%), according to the SCL-90. At previous measurements, the distribution was rather skewed, which hampered a sensitivity analysis.

For the OQ, the sensitivity (correct positives) was 0.93 and specificity (correct negatives) 0.69. The sensitivity for the HoNOS was 0.83 and 0.72 for specificity. The OQ shows an area under the curve (AUC) of 0.82; for the HoNOS the AUC was 0.77. The values found for both the OQ and the HoNOS were good predictors of the results of the SCL-90.

## 5. Discussion

The combination of the observer-rating scale HoNOS and the self-rating scale OQ is a promising substitute for the SCL-90 when using ROM in psychiatry. We give

## Chapter 3 | Validation of Two Outcome Measures

some arguments for the combination and are willing to make some recommendations for use in clinical practice.

1. Firstly, application of the two measures takes approximately a quarter of an hour; much less time is needed than the SCL-90.
2. Secondly, the psychometric analysis has shown that the two measures are of sufficient quality. On the one hand, the reliabilities of both measures are good and scores correspond sufficiently with those of the SCL-90, which contributes to the convergent validity. On the other hand, scores for the various groups of patients vary (discriminant validity). It was assumed that the SCL-90 ascertained the more subjective perceptions of the patient, whilst the subscales of the OQ and the HoNOS together could be seen as more objective measures of psychological problems. Comparison with the SCL-90 as the gold standard produces good results and shows that the criterion validity of the two measures is adequate.
3. Thirdly, both measures are sufficiently sensitive to change, as is apparent from the CSC values results. It is true that the HoNOS shows a much lower CSC than the other two outcome measures, but 8% improvement between measurements 1 and 3 is acceptable. Parabiaghi et al. (2005) found 5.6% and regarded this as adequate. In our sample, 7.4% show an improvement between T1 and T2, but only 3.4% show a clinically significant improvement as well. It is sensible to use these figures (RC, CSC, and the number of individuals surpassing the cut-off) in combination when assessing an improvement and to realize that CSC is a stringent measure that produces low percentages. The HoNOS is also often used as an observer rating scale for patients with more severe psychiatric disorders. As is to be expected, a much higher change value is awkward; the instrument consists of only 12 items and the criterion for improvement is quite stringent (no single item has a value of 2 or more), as compared to the OQ and the SCL-90. In practical terms, one must take into account the fact that this instrument shows only small changes that, however, could be significant.  
Furthermore, it is recommended to use the total HoNOS score when carrying out a study because the subscales are not sufficiently reliable. The subscales of the OQ can be interpreted independently in terms of improvement and deterioration.
4. Fourth, together the two outcome measures represent the most important domains, functionally as well as symptomatically, to assess the **effects of treatment. Combining “clinical judgment” (HoNOS) with a self-**

## Validation of Two Outcome Measures | Chapter 3

report on subjective well-being (OQ) also has a practical advantage in that the agreement or non-agreement of the various outcomes can be **informative of the patient's situation. Information obtained from both** instruments can complement each other. Mulder et al. (2010) reported that in cases of severe psychiatric disorders, the HoNOS and the Manchester Short Assessment of Quality of Life (MANSA, a self-report measure) are correlated but also complementary: this is because the HoNOS reflects the perspective of the clinician, whereas the MANSA reflects that of the patient.

Using only the SCL-90 for ROM, as is often the case, is thus less informative. Additionally, the SCL-90 was explicitly designed for (epidemiological) research and is therefore less targeted to clinical practice. Furthermore, we found that patients with multiple severe diagnoses show the greatest improvement using the HoNOS. They begin with the worst scores and can profit the most from treatment.

One of the disadvantages of this study is that the sample is not representative of all psychiatric diagnoses. Future studies must validate the measures for samples in which more severe psychiatric disorders occur, for example, in a clinical setting, bearing in mind that the drop-out rate after the first measurement is high.

## Chapter 3 | Validation of Two Outcome Measures

### References

Aarsse R. Betekenis van cliënttevredenheid als indicator voor kwaliteit van zorg. Amsterdam: Thela Publishers 2003.

Aartsen MJ, Spitsbaard AK, Van Baarsen C, Dhondt, ADF, Mascini M, Nefs A, et al. Een multicenterstudie naar betrouwbaarheid, validiteit en gevoeligheid voor verandering van de HoNOS65+ binnen de ouderenpsychiatrie. Tijdschr psychiatr 2010; 52(8): 543-553.

Arindell WA, Ettema JHM. SCL-90: Handleiding bij een multidimensionele psychopathologie-indicator. Lisse: Swets & Zeitlinger 1986/2003.

Beurs E de, Den Hollander-Gijsman M, Buwalda V, Trijsburg W & Zitman F. De Outcome Questionnaire (OQ-45): psychodiagnostisch gereedschap. De Psycholoog 2005; 40: 393-400.

Broersma TW & Sytema S. Implementatie van het meetinstrument HoNOS65+. Onderzoek op een afdeling ouderenpsychiatrie. Tijdschr psychiatr 2008; 50(2): 77-82.

Chapman JE. Reliability and validity of the Progress Questionnaire: an adaptation of the Outcome Questionnaire. Drexel PhD Thesis; 2003.

Derogatis LR. The SCL-90 Manual: Scoring, administration and procedures for the SCL-90. Baltimore, MD John Hopkins University School of Medicine 1977.

Doerfler LA, Addis, ME & Moran PW. Evaluating mental health outcomes in an inpatient setting: convergent and divergent validity of the OQ-45 and Basis-32. J Behav Health Serv Res 2002; 29(4): 394-403.

Hermann RC. Quality Assessment and Improvement in a Changing Healthcare System. In: Hermann, R.C., editor. Improving Mental Healthcare. A Guide to Measurement-Based Quality Improvement. Arlington: American Psychiatric Publishing Inc.; 2005.

Holi MM, Martunen M & Aalberg V. Comparison of the GHQ-36, the GHQ-12 and the SCL-90 as psychiatric screening instruments in the Finnish population. Nord J Psychiatry 2003; 57(3): 233-238.

## Validation of Two Outcome Measures | Chapter 3

Hunter R, Cameron R & Norrie J. Using patient reported outcomes in schizofrenia: the Scottish schizofrenia outcomes study. *Psychiatr Serv* 2009; 60(2): 240-245.

Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991; 59: 12-19.

Jong K de, Nugter MA, Polak MG, Wagenborg JEA, Spinhoven P, Heiser W. The Outcome Questionnaire (OQ-45) in a Dutch Population: A Cross-Cultural Validation. *Clin Psychol Psychother* 2007; 14: 288-301.

Koeter M. Validity of the GHQ and SCL anxiety and depression scales: a comparative study. *J Affect Disord* 1992; 24: 271-280.

Lambert MJ, Burlingame GM, Umpress V, Hansen NB, Vermeersch DA, Clouse GC, et al. The Reliability and Validity of the Outcome Questionnaire. *Clin Psychol Psychother* 1996; 3(4): 249-258.

Lambert M, Hansen NB & Finch AE. Client focused research: using client outcome data to enhance treatment effects. *J Consult Clin Psychol* 2001; 69(2): 159-172.

Mulder CL, Staring ABP, Loos J, Buwalda VJA, Kuijpers D, Sytema S et al. De Health of the Nation Outcome Scales (HoNOS) als '**routine outcome assessment**'. *Tijdschr psychiatr* 2004; 46(5): 274-284.

Mulder CL, Van der Gaag M, Bruggeman R, Cahn W, Delespaul PAE, Dries P, et al. Routine outcome monitoring voor patiënten met ernstige psychiatrische aandoeningen; een consensusdocument. *Tijdschr psychiatr* 2010; 52(3): 169-179.

Oiesvold T, Bakkejord T, Sexton, AJ. Concurrent validity of HoNOS compared with a patient derived measure (SCL-90-R) in outpatients' clinics. ***Psychiatry Res* 2011; 187: 297-300.**

Orrell M, Yard P, Handysides J, Schapira R. Validity and reliability of the Health of the Nation Outcome Scales in psychiatric patients in the community. *Br J Psychiatr* 1999; 174: 409-12.

## Chapter 3 | Validation of Two Outcome Measures

Parabiaghi A, Barbato A, D'Avanzo P, Erlicher A, Lora A. Assessing reliable and clinically significant change on health of the nation outcome scales: methods for displaying longitudinal data. *Aust N Z J Psychiatry* 2005; 39: 719-725.

Sharma VK, Wilkinson G, Fear S. Health of the nation outcome scales: a case study in general psychiatry. *Br J Psychiatry* 1999; 174: 395-398.

Slade M, Beck A, Bindman T, Thornicroft G, Wright S. Routine clinical outcome measures for clients with severe mental illness: CANSAS and HoNOS. *Br J Psychiatry* 1999; 174: 404-408.

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al.. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psych* 1998; 59:22-33

Thornicroft G, Bebbington P, Leff J. Outcomes for long-term clients one year after discharge from a psychiatric hospital. *Psychiatr Serv* 2005; 56(11): 1416-1422.

Trauer T, Callaly T, Hantz P, Little J, Shields R, Smith J. Health of the Nation Outcome Scales. Results of the Victorian field trial. *Br J Psychiatry* 1999; 174: 380-388.

Umphress VJ, Lambert MJ, Smart DJ. Concurrent and construct validity of the outcome questionnaire. *J Psychol Educ Assess* 1997; 15: 40-55.

Vermeersch DA, Lambert MJ, Burlingame GM. Outcome Questionnaire: item Sensitivity to Change. *J Pers Assess* 2000; 74(2): 242-261.

Vliet IM van, Beurs E de. Het Mini internationaal neuropsychiatrisch interview. Een kort gestructureerd diagnostisch psychiatrisch interview voor DSM-IV en ICD-10 stoornissen. *Tijdschr psychiatr* 2007; 49(6): 393-307.

Wilson, SA, Becker, LA, Tinker, RH. Fifteen month follow-up of eye movement desensitization and reprocessing (EMDR) treatment for posttraumatic stress disorder and psychological trauma. *J Consult Clin Psychol* 1997; 65(6): 1047-1056.

## Validation of Two Outcome Measures | Chapter 3

Wing, JK, Beevor, AS, Curtis, RH, Park, BG, Hadden, S, Burns, H. Health of the nations outcome scales (HoNOS): research and development. *Br J Psychiatry* 1998; 172: 11-18.

Dissertation Series