*Clinical Psychology and Psychotherapy*
*Clin. Psychol. Psychother.* **24**, 72–81 (2017)
Published online 26 October 2015 in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/cpp.1979

# *Cut-off Scores and Clinical Change Indices for the Dutch Outcome Questionnaire (OQ-45) in a Large Sample of Normal and Several Psychotherapeutic Populations*

**Reinier Timman,[1]* Kim de Jong[2] and Nita de Neve-Enthoven[2]**
[1] *Department of Psychiatry, Medical Psychology and Psychotherapy Section, Erasmus Medical Centre Rotterdam, Rotterdam, The Netherlands*
[2] *Leiden University, Institute of Psychology, Clinical Psychology Unit, Leiden, The Netherlands*

**The Outcome Questionnaire-45 (OQ-45; Lambert *et al.*, 2004) has been designed for frequent assessment of a patient's functioning during the course of psychotherapy and has become one of the most frequently used outcome measures in the Netherlands. The OQ-45 was originally normed on outpatients in secondary care only, but is applied in a wide variety of patient populations. As such, it has become increasingly important to provide cut-off scores with the normal population, as well as between different patient populations. The present large-scale Dutch study aims to provide cut-off scores between several populations. Data were collected from the normal population (*n* = 1810) and patients in five different treatment settings: outpatient primary care (*n* = 1581), outpatient secondary care (*n* = 9433), private practice (*n* = 457), day patient (*n* = 481) and inpatient therapies (*n* = 485), a total of more than 14.000 administrations. Reliable change indices and cut-off scores were calculated using the method of Jacobson and Truax (1991). The reliable change index for the patient population was calculated as 18 and the cut-off between the normal and patient population as 56. Sensitivity, specificity and area under the curves of cut-off scores between the normal population and the treatment settings were satisfactory and generally higher than 0.80. Between the patient populations, these measures were generally too low for strict use. The OQ-45 total score can satisfactorily discriminate between the normal and patient populations. For assignment to specific treatment types, the OQ-45 may help, but its use is somewhat limited in practice.**

**Key Practitioner Message:**

- The Dutch OQ-45 has satisfactory levels of reliability, sensitivity, specificity and area under the curve.
- The new overall cut-off score for normal function for the Dutch OQ-45 is 56 and the new reliable change index is 18.
- Cut-off scores for several therapeutic treatments are provided.

**Keywords:** OQ-45, Cut-off, Sensitivity, Specificity

## INTRODUCTION

The self-report Outcome Questionnaire-45 (OQ-45) was designed as a brief measure of patient progress in a therapeutic setting. The instrument's ability to reflect clinical change during interventions, such as psychotherapy, has made it a useful tool for high frequent routine outcome monitoring (ROM) of psychotherapy, also referred to as tracking. To determine whether a patient's progression is positive or negative, the individual score of a patient can be compared with the score that would be expected for this patient (Finch, Lambert, & Schaalje, 2001). Tracking, and the provision of feedback to therapists and sometimes also to patients, is found to reduce treatment failure and prevents overtreatment (Lambert, 2005; Shimokawa, Lambert, & Smart, 2010). Feedback based on the OQ-45 supplemented with the use of clinical support tools (CSTs) has been shown to be especially effective in improving treatment outcomes (Shimokawa *et al.*, 2010). CSTs can be described as empirically based problem-solving strategies, helping therapists to focus on important factors that can influence the outcome of psychotherapy (Probst *et al.*, 2013). The OQ-45 has been used in a wide range of clinician settings, ranging from counselling centres

*Correspondence to: Reinier Timman, Erasmus Medical Centre Rotterdam, Department of Psychiatry, Medical Psychology and Psychotherapy Section, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands. E-mail: r.timman@erasmusmc.nl

(e.g., Lambert *et al.*, 2001) to inpatient settings (e.g., Simon *et al.*, 2012; Probst *et al.*, 2013).

There are several ways to determine criteria for sufficient or insufficient change in patient progress. The OQ-45 and many other questionnaires make use of cut-off scores, which indicate whether the person's reported psychological distress is of clinical significance; a person scoring above the cut-off score is considered to experience more psychological distress than a person from the normal population. Additional methods focus on the amount of change a person has made over time, to determine rules for evaluating treatment progress. For instance, Lutz *et al.* (2013) describe a method based on using sudden gains and sudden losses between sessions. This can be determined by a 25% gain or loss relative to the previous session, or by a significant gain or loss of two or three sessions relative to the previous two or three sessions. Another frequently used method is the use of expected recovery curves ( Finch *et al.*, 2001; Probst *et al.*, in press; Newnham, Hooke, & Page, 2010). Expected recovery curves are based empirically on norm groups of patients with characteristics comparable with the patient in question. Extreme positive or negative deviations are indicative of treatment success. These methods can be built into a routine outcome monitoring system, but as these methods are dependent on change of individuals' scores, they cannot be applied to determine a general number of points for change. Moreover, cut-off scores have the advantage over change-based criteria that they provide a clear criterion for patient recovery.

Another relevant concept in ROM is whether a change is large enough not to be caused by measurement error. A reliable change index (RCI) can be calculated, a reported change in clinical symptoms equal to or larger than this RCI is expected not to be due to chance fluctuations and thus to reflect a clinically significant change, in other words, a change greater than the RCI implies that the change will 95% certain not be due to measurement error (Jacobson & Truax, 1991). Tingey, Lambert, Burlingame, and Hansen (1996) elaborated the RCI concept. They state that there is no single RCI for a given questionnaire, but that the RCI is dependent on the severity of the population. Patient and the normal populations can be ordered on a continuum, and RCIs must be calculated considering the place of the population on that continuum. Schauenburg and Strack (1999) also recognize the need to identify multiple patient populations and they advocate to determine cut-off points and other indices in non-English-speaking countries. Properly determined cut-off scores and RCIs can make an important contribution to the decision-making process for reduction of treatment frequency and ultimately treatment termination. However, the decision for termination of therapy also depends on many other characteristics that should be considered by therapist and patient. In contrast to other outcome monitoring instruments that require considerable time to fill out or target only specific patient populations, the OQ-45 can be completed in approximately 10 min and can be used in a broad range of patients with a variety of psychiatric diagnoses. These advantages have made the instrument one of the most used outcome instruments in the fields of psychology and psychiatry worldwide. High levels of reliability have been reported (Cronbach's α 0.93), and concurrent validity estimates were found to be excellent in a number of studies (Durham *et al.*, 2002).

The total score for the OQ-45 describes the general mental health factor. The higher the reported total score, the larger the distress experienced by the respondent. Additionally, the OQ-45 uses three subscales to measure three domains of functioning. First, the Symptom Distress (SD) scale measures psychological features characteristic for mental disorders. Second, the Interpersonal Relations (IR) scale measures problems in and satisfaction with relationships with partner, family and friends. Third, the Social Role Performance (SR) scale measures the respondent's perceived functioning in social roles such as in work, school or leisure. While examining the factor structure of the instrument for the Dutch population, De Jong *et al.* (2007) found a fourth subscale next to the three factor solution, the Anxiety and Social Distress (ASD) scale, which might be of interest to therapists that specialize in anxiety or psychosomatic disorders.

The OQ-45 has up to now been translated into 32 languages. Studies on the psychometric properties of the translated questionnaires report differences in reported cut-off scores and RCIs between countries. For example, Amble *et al.* (2014) studied the psychometric properties of the Norwegian version of the OQ-45 and compared their findings with the US, German (Moessner, Gallas, Haug, & Kordy, 2011), Italian (Chiappelli *et al.*, 2008; Lo Coco *et al.*, 2008), Swedish (Wennberg, Philips, & De Jong, 2009), Dutch (De Jong *et al.*, 2007) and Chinese (Qin & Hu, 2008) studies. Reported means, cut-off scores and RCIs of clinical and non-clinical populations varied substantially internationally. These differences can be due to cultural differences or semantic differences of the translation. Cultural values can influence the presentation and experience of psychological distress (Bhugra *et al.*, 2011). Additionally, differences in OQ-45 mean scores might reflect true differences in well-being between countries, for example due to the economic situation in the area under study (Amble *et al.*, 2014). De Jong *et al.* (2007) state that even among Western countries, differences in normative scores are a regularly observed phenomenon.

Differences can also depend on the therapy groups under study. For instance, inpatients are expected to have higher baseline scores than outpatients. Consequently, a cut-off score between inpatients and the normal population will be higher than a cut-off score between outpatients and the normal population or inpatients and outpatients. A concern in this respect is the different composition of norm groups across countries. Amble *et al*. (2014) reported that the clinical samples in the countries included in their study differed substantially in sample size and characteristics, which might have influenced the mean scores of those groups as well as the overall cut-offs for normal functioning. Characteristics of clinical samples depend on the nationwide organization of the mental healthcare system and the attitude of inhabitants towards psychological treatment. For example, in Argentina, psychotherapy is viewed as a useful tool in the process of the general development of a person, and visiting a therapist is therefore commonplace (Digiuni, Jones, & Camic, 2013). The scores of these 'patients' on outcome measurement tools might however have a substantial influence on the distribution of scores and the associated means of psychological distress of the patient population. Amble *et al*. (2014) conclude that since cut-off scores, norms and RCIs are not transportable across countries, they should be separately determined for each country if one aims to use the OQ-45 scores as a valid indicator of psychological distress.

A relatively large amount of research on the psychometric properties of the Dutch OQ-45 has been conducted (De Beurs *et al*., 2005; De Jong *et al*., 2007) and supported that the Dutch OQ-45 had good psychometric properties. However, the cut-off scores for normal function and RCI were based on outpatients in secondary care only, even though the instrument was intended for use in a broader range of treatment settings. This may explain why a considerably higher cut-off score has been reported for the original US version (63) (Lambert *et al*., 2004) than the Dutch version, which De Jong, Nugter, Lambert, and Burlingame (2009) reported to be 55, and De Beurs *et al*. (2005) reported to be 56 for men and 57 for women.

Currently, multiple transitions in the organization of the Dutch mental healthcare system are being made. Protocols prescribe to predominantly treat mental health problems in primary care. Primary mental health care is executed by the general practitioner, psychologists and psychotherapists, and is accessible for any patient without referral. The general practitioner functions as a gatekeeper and is responsible for referral of the patient to other professionals when more specialized care is required (Thomson, Osborn, Squires, & Jun, 2013). Secondary mental health care includes specialized mental healthcare institutions, and independent psychologists, psychotherapists and psychiatrists. When professionals intend to use the OQ-45 for triage purposes, or to evaluate treatment progress in different treatment settings, a more elaborate study on cut-off scores between the normal population and various patient populations is needed. Accurate cut-off scores and associated RCIs are necessary for allocation of patients to the most suitable treatment form and the tracking of treatment progress in different patient populations. Additionally, cut-off scores between populations can aid referral to lower intensity (and lower costs) treatment forms, when possible. Finally, an overall cut-off score between a balanced representation of the patient populations (based on the proportion of patients per treatment type) and the non-clinical population is necessary, to determine when the level of functioning of a patient is comparable with someone who is not experiencing psychological distress.

The present study aims to determine the OQ-45 cut-off scores and RCIs for the normal and various patient populations in the Netherlands. Additionally, a new overall cut-off score will be computed, based on a stratified sample from five patient populations. Data were collected in the general population, primary and secondary outpatients, private practices, day patients and inpatients, together comprising more than 14.000 administrations.

## METHOD

### *Subjects*

The study included six samples: a sample from the normal population and five patient samples. All patient data were collected at the start of treatment.

The normal population consisted of 1810 subjects and was an aggregation of two normative samples. A community sample of 1000 individuals representative of the Dutch adult population regarding sex, age, social economic status and education was collected by TNS-NIPO. TNS-NIPO is a renowned polling and marketing institute in the Netherlands. They have a large pool of individuals who receive a small incentive for participation in a study or poll. Subjects were not screened for active treatment. Data of a community sample of 810 individuals that were collected in the validation study by De Jong *et al*. (2007) were added to this sample. These individuals were selected from business and non-profit organization samples, and from the phone directory. Twenty-four subjects who received treatment for psychological or psychiatric problems were excluded from the sample. A difference in OQ-45 score was observed between the two subsamples. The TNS-NIPO group had a mean total OQ-45 score of 42.6 ($SD = 18.9$), and the sample by De Jong *et al.* had a mean of 38.7 ($SD = 16.1$; Cohen's $d = 0.22$; $p < 0.001$).

The private practice sample of 457 subjects was collected in the monitoring research program of Erasmus MC, department of Medical Psychology and Psychotherapy (De Jong *et al*., 2014). Although private practices treat primary as well as secondary care patients, no information on this distinction was available.

The two outpatient samples of 1581 (outpatient primary care) and 9433 (outpatient secondary care) individuals were collected in the monitoring programs of Indigo

(Castricum), Amici (Noord-Holland), psychotherapeutic centres GGZ Dijk & Duin and GGZ Noord-Holland-Noord (Noord-Holland), De Viersprong (Halsteren), PsyQ (The Hague) and the previously mentioned monitoring research program (De Jong *et al*., 2014).

The samples of 481 day patients and 484 inpatients, respectively, were collected in the SCEPTRE research program carried out by De Viersprong (Soeteman, Hakkaart-van Roijen, Verheul, & Busschbach, 2008) and at GGZ-Centraal (Amersfoort). These were long-term psychotherapeutic programs for personality disorders. Day patients typically receive treatment programs similar to inpatient programs, but do not stay overnight, although some programs allow for occasional overnight stays. Inpatient treatments included at least four overnight stays in the treatment centre.

For the creation of an overall stratified patient group data from the Dutch central statistics agency (Centraal Bureau voor de Statistiek; CBS) was used to determine the proportions of the patient groups in Dutch society. CBS reports no distinction between primary versus secondary care treatments, and between day patients and inpatients. Therefore, we assumed a fifty-fifty distribution for these subgroups. CBS reports for the proportion of therapies in private practices and institutes as 16% and 84%. The proportions of inpatients and outpatient therapies are 6% and 94%. Assuming no interaction, it can be inferred that 79.2% of all patients are outpatients in an institute setting, 14.8% are outpatients in private practices, 5.1% are inpatients in an institute and 1.0% should be private practice inpatients, although in practice, there are no inpatients in private practices. We randomly selected five samples with the determined proportions from each of the patient populations. This resulted in five data sets used to determine the overall cut-off score for normal functioning, consisting of 1224 outpatients in primary care, 1224 outpatients in secondary care, 79 day patients, 79 inpatients and 457 patients in private practices, a total of 3063 patients.

### Instrument

The OQ-45 consists of 45 items that are scored on a 5-point Likert scale ranging from 0 (*never*) to 4 (*almost always*). Respondents are asked to fill out the questionnaire based on how they have felt over the past week. The questionnaire can generally be administered within 5 to 10 min. The present study used the validated Dutch translation of the OQ-45.

The total score for the OQ-45 (range: 0–180) can be calculated by adding up the scores on all 45 items. The Symptom Distress scale (SD) has 25 items and a range of 0–100. The IR scale (IR) has 11 items and a range of 0–44. The SR scale (SR) has nine items and a range of 0–36. The ASD scale consists of 13 items with scores ranging 0–52. In prior research, Cronbach's α values for internal consistency were found to be sufficient for most subscales and the total scale. Reliability values for the SR scale have proven to be unsatisfactory in international research, mainly due to the poor functioning of item 14 'I work/study too much' (e.g., De Jong *et al*., 2007; Lambert *et al*., 2004).

### Data Analysis

All statistical analyses were performed using IBM SPSS Statistics (version 21, IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY, USA). Differences in continuous variables (age and OQ-45 scores) between the study samples were analysed using one-way analysis of variance. For pairwise comparisons, no correction has been applied for multiple testing, as no hypotheses have been stated regarding the equivalence or difference of the samples. Differences in proportions were analysed with Yates-corrected chi-square ($\chi^2$) tests. All tests of difference were two-tailed against $p < 0.05$.

#### *Reliable Change Index, Reliability*

As measures of reliability, we calculated Cronbach's alphas and standard errors of measurements. The RCIs were calculated with the method described by Jacobson and Truax (1991). Cronbach's alpha was used as the measure of reliability ($r_{xx}$) in the RCI formula (Equation 1).

$$RC = 1.96\ \sigma\ \sqrt{2(1 - r_{xx})} \tag{1}$$

#### *Cut-off Scores*

Jacobson and Truax (1991) also described a method for the determination of cut-off scores. As we had information of several populations, we chose the method of determining the score where the proportion of cases in population 1 below that score equals the proportion of cases above that score in population 2. This cut-off point (c) is determined by Equation 2, where $x_{1,}$ $x_2$ are the means and $\sigma_1$, $\sigma_2$ are the standard deviations of the respective populations.

$$c = \frac{\sigma_1 \overline{x}_2 + \sigma_2 \overline{x}_1}{\sigma_1 + \sigma_2} \tag{2}$$

This method assumes normal distributions of the outcome variable in the populations at hand, in particular with respect to skewness. Shapiro–Wilks tests and the procedure described by Tabachnick and Fidell (1996) were applied to check whether this normality assumption was violated, and if so, in what way.

### *Sensitivity, Specificity, Overall Accuracy and Area Under the Curve*

Sensitivity is the proportion of correctly classified patients in a treatment setting (or normal population) of all actual patients in that therapy category. Specificity is the proportion of patients not categorized in that treatment setting (or normal population), which are correctly identified as such. Sensitivity and specificity are combined in an overall accuracy measure and the area under the curve (AUC, Hanley & McNeil, 1982). For calculation of the sensitivity, specificity and overall accuracy, we randomly drew two halves of the participants from each sample and calculated all cut-off measures in the first sample. Subsequently, we calculated sensitivity, specificity and overall accuracy in the second (replication) sample. In this way, we avoid artificially inflated classification estimates by criterion contamination.

## RESULTS

### *Demographic Characteristics*

Generally, more women (61%) than men were included in the study (Table 1). In particular, most women were included in day patient and inpatient treatments. Our normal sample also included more women than men, which is in concordance with the higher proportion of women in the patient populations.

Age differed significantly between most populations ($p < 0.001$), except for private practice patients and primary care outpatients ($p = 0.38$), and for part-time and full-time inpatients ($p = 0.26$). The normal population included fewer young participants than the treatment samples.

### *Group Differences in Outcome Questionnaire-45 Scores*

More severe patient groups tended to have significantly higher OQ-45 total scores ($p < 0.001$) than other patient groups and the normal sample, with the exception of day patients and inpatients ($p = 0.069$).

The age group 20–39 scored highest on the total OQ-45 score ($p < 0.001$), whilst age group $\geq 60$ scored lowest ($p < 0.001$). No significant differences were found between age groups $< 20$ and 40–59 ($p = 0.27$). Women scored higher than men ($p < 0.001$). Private practice patients had significant more years of education than the other groups ($p < 0.001$, overall effect size 0.52).

### *Reliability*

Internal consistency values as calculated with Cronbach's alpha were found to be good for the total OQ-45 scale and SD subscale in the normal population as well as the patient populations (range 0.89–0.93, Table 2). The reliability values for IR were moderate (range 0.72–0.78), and the values for SR were almost all below 0.70 and must therefore be considered poor (Nunnally, 1978). Reliability values for the ASD were good (range 0.80–0.86). These findings indicated that the calculated cut-off scores for the subscale IR were less reliable. The RCIs of the total scale are 18 in the overall patient sample and ranged from 14.4 in the normal population to 19.6 in the outpatient secondary care group.

The normal and secondary care outpatient population group had distributions that significantly deviated from normality (data not shown). This could argue for a non-parametric approach. In particular, the distribution in the normal population was skewed in such a way that relatively much subjects had low scores and few had high

Table 1. Sample characteristics and Outcome Questionnaire-45 total scores

| | Normal population | Private practice | Outpatient primary care | Outpatient secondary care | Part-time inpatient | Full-time inpatient | Total sample |
|---|---|---|---|---|---|---|---|
| *n* (%) | | | | | | | |
| Total | 1810 (13) | 457 (3) | 1581 (11) | 9433 (66) | 481 (3) | 484 (3) | 14 246 (100) |
| Male | 769 (42) | 163 (36) | 537 (34) | 3542 (41) | 128 (27) | 145 (30) | 5284 (39) |
| Female | 1041 (58) | 294 (64) | 1032 (66) | 5085 (59) | 353 (73) | 339 (70) | 8144 (61) |
| Mean age (*SD*) | | | | | | | |
| Total | 46 (16) | 40 (12) | 40 (13) | 37 (12) | 31 (10) | 30 (10) | 38 (13) |
| Male | 49 (17) | 42 (12) | 41 (12) | 38 (12) | 33 (10) | 32 (10) | 40 (14) |
| Female | 44 (15) | 39 (12) | 40 (13) | 36 (12) | 30 (10) | 29 (9) | 37 (13) |
| Mean years of education *(SD)* | | | | | | | |
| Total | 14.5 (3.4) | 16.1 (3.2) | * | 14.5 (3.3) | 14.3 (3.1) | 14.6 (3.0) | 14.6 (3.3) |
| Male | 14.7 (3.6) | 16.3 (3.2) | | 14.8 (3.3) | 14.0 (3.1) | 14.8 (3.0) | 14.8 (3.4) |
| Female | 14.3 (3.3) | 16.1 (3.2) | | 14.3 (3.3) | 14.4 (3.1) | 14.5 (2.9) | 14.5 (3.2) |
| Mean OQ total score (*SD*) | | | | | | | |
| Total | 40.9 (17.8) | 67.4 (22.1) | 73.4 (21.5) | 77.9 (24.7) | 89.6 (20.9) | 92.3 (20.2) | 73.2 (26.7) |
| Male | 40.0 (17.7) | 64.8 (22.4) | 72.9 (22.3) | 75.8 (25.2) | 87.6 (18.9) | 93.3 (17.6) | 70.7 (27.1) |
| Female | 41.5 (17.9) | 68.8 (21.9) | 73.6 (21.0) | 79.2 (24.2) | 90.3 (21.5) | 91.9 (21.2) | 74.3 (26.4) |

*No data available

Table 2. Means, standard deviations, Cronbach's alphas and of the Outcome Questionnaire-45 total scale and subscales

| | Normal population | Private practice | Outpatient primary care | Outpatient secondary care | Part-time outpatient | Full-time outpatient | Overall* patient |
|---|---|---|---|---|---|---|---|
| Mean (*SD*) | | | | | | | |
| Total scale [range 0–180] | 40.9 (17.8) | 67.4 (22.1) | 73.4 (21.5) | 77.9 (24.7) | 89.6 (20.9) | 92.3 (20.2) | 75.1 (23.6) |
| Symptom distress [0–100] | 23.8 (11.5) | 40.5 (14.4) | 45.3 (13.9) | 48.2 (16.1) | 54.1 (13.4) | 56.2 (13.5) | 46.2 (15.3) |
| Interpersonal relations [0–44] | 9.0 (5.0) | 14.5 (6.1) | 15.1 (6.1) | 16.3 (6.9) | 20.3 (6.5) | 21.0 (6.1) | 15.8 (6.6) |
| Social role [0–36] | 8.1 (3.9) | 12.4 (5.1) | 13.1 (5.2) | 13.1 (5.5) | 15.0 (4.9) | 15.1 (5.0) | 13.1 (5.4) |
| Anxiety and social distress [0–52] | 14.4 (6.8) | 21.5 (8.0) | 25.1 (7.9) | 25.6 (9.3) | 27.2 (8.1) | 27.8 (7.9) | 24.9 (8.7) |
| Cronbach's alpha | | | | | | | |
| Total scale | 0.92 | 0.93 | 0.92 | 0.92 | 0.90 | 0.90 | 0.92 |
| Symptom distress | 0.89 | 0.92 | 0.90 | 0.91 | 0.89 | 0.89 | 0.91 |
| Interpersonal relations | 0.75 | 0.77 | 0.76 | 0.78 | 0.75 | 0.72 | 0.78 |
| Social role | 0.58 | 0.72 | 0.69 | 0.66 | 0.59 | 0.59 | 0.68 |
| Anxiety and social distress | 0.82 | 0.83 | 0.81 | 0.85 | 0.80 | 0.80 | 0.84 |
| Standard error of measurement | | | | | | | |
| Total scale | 5.18 | 5.92 | 6.26 | 6.82 | 6.66 | 6.48 | 6.51 |
| Symptom distress | 3.73 | 4.14 | 4.44 | 4.72 | 4.55 | 4.53 | 4.55 |
| Interpersonal relations | 2.50 | 2.91 | 2.96 | 3.25 | 3.27 | 3.21 | 3.11 |
| Social role | 2.52 | 2.69 | 2.91 | 3.21 | 3.17 | 3.18 | 3.02 |
| Anxiety and social distress | 2.89 | 3.28 | 3.47 | 3.63 | 3.61 | 3.58 | 3.52 |
| Reliable Change Index | | | | | | | |
| Total scale | 14.4 | 16.4 | 17.3 | 18.9 | 18.5 | 18.0 | 18.0 |
| Symptom distress | 10.3 | 11.5 | 12.3 | 13.1 | 12.6 | 12.5 | 12.6 |
| Interpersonal relations | 6.9 | 8.1 | 8.2 | 9.0 | 9.1 | 8.9 | 8.6 |
| Social role | 7.0 | 7.5 | 8.1 | 8.9 | 8.8 | 8.8 | 8.4 |
| Anxiety and social distress | 8.0 | 9.1 | 9.6 | 10.1 | 10.0 | 9.9 | 9.8 |

*Mean values of five stratified samples, $n = 3063$.

scores. This skewness can be caused by a bottom effect; normal individuals obviously are expected to score low on the OQ-45. The deviation in the primary outpatient group was caused by a high kurtosis, not by skewness. For the determination of cut-off scores, the overlap of the distributions is critical, and the left-sided non-normality of the normal population does not concern an overlapping part of the distribution. For this reason, we argue that there was no objection for application of the method for calculating cut-off scores suggested by Jacobson and Truax (1991).

The cut-off scores between the normal population and the several patient populations that were calculated in the first stratified random half sample are presented in Table 3. This table also reports the classification accuracies that were calculated in the replication sample. Regarding the total OQ-45 scale, sensitivity, specificity, overall accuracy and AUC indices were acceptable for discrimination between the normal population and the private practice therapies. These measures were good for discriminating between the normal population and patient populations (Table 3). The same holds for the SD subscale, and to some extent to the ASD subscale. For the IR and SR subscales, these measures must be considered poor in discriminating between the normal population and private practice and primary and secondary outpatients, and acceptable to good for discriminating between the normal and patient populations.

The overall cut-off score between the normal and patient population is 56. Cut-off scores and figures for discrimination between the therapy forms are given in Table 4. Differences in cut-off scores between men and women were up to three points. For means of conciseness, the AUCs are not reported, as these measures are virtually equal to the overall accuracies. Generally, these measures must be considered poor (< 0.70). Only the distinction between the private practice and inpatient groups can be considered moderate (± 0.70).

## DISCUSSION

In this study, we determined the cut-off scores between, and RCIs within the normal population and various psychotherapeutic settings for the OQ-45 in the Netherlands using the methods described by Jacobson and Truax (1991), Tingey *et al*. (1996), Schauenburg and Strack (1999) and Hanley and McNeil (1982). These psychotherapeutic settings included private practice therapies, primary and secondary outpatient care, day patient and inpatient therapies. We also constructed a stratified, combined patient population for the calculation of overall cut-off scores.

As expected, patients with the highest OQ-45 scores, reflecting more psychological problems, were treated in the most intensive treatments (day patients and inpatients).

Table 3. Cut-off scores, sensitivity, specificity, overall accuracy and area under the curve for the Outcome Questionnaire-45 total scale and subscales for discrimination between the normal and patient populations

| | Private practice | Outpatient primary care | Outpatient secondary care | Part-time inpatient | Full-time inpatient | Overall patient* |
|---|---|---|---|---|---|---|
| Cut-off | | | | | | |
| Total scale | 52.5 | 55.5 | 56.4 | 63.6 | 64.7 | 55.6 |
| Men | 51.2 | 55.1 | 55.2 | 63.5 | 67.9 | 53.9 |
| Women | 53.8 | 56.1 | 57.2 | 62.8 | 64.3 | 56.0 |
| Symptom distress | 31.1 | 33.5 | 33.9 | 37.9 | 38.6 | 33.4 |
| Interpersonal relations | 11.5 | 11.7 | 12.2 | 14.1 | 14.5 | 11.9 |
| Social role | 9.9 | 10.2 | 10.2 | 11.1 | 11.0 | 10.2 |
| Anxiety and social distress | 17.6 | 19.4 | 19.1 | 20.2 | 20.7 | 19.0 |
| Sensitivity | | | | | | |
| Total scale | 0.72 | 0.80 | 0.80 | 0.86 | 0.90 | 0.78 |
| Symptom distress | 0.72 | 0.79 | 0.81 | 0.88 | 0.90 | 0.78 |
| Interpersonal relations | 0.69 | 0.71 | 0.69 | 0.76 | 0.83 | 0.72 |
| Social role | 0.72 | 0.66 | 0.65 | 0.76 | 0.75 | 0.66 |
| Anxiety and social distress | 0.66 | 0.74 | 0.74 | 0.82 | 0.82 | 0.74 |
| Specificity | | | | | | |
| Total scale | 0.78 | 0.83 | 0.83 | 0.89 | 0.90 | 0.82 |
| Symptom distress | 0.77 | 0.81 | 0.81 | 0.88 | 0.90 | 0.81 |
| Interpersonal relations | 0.73 | 0.73 | 0.80 | 0.87 | 0.87 | 0.73 |
| Social role | 0.68 | 0.64 | 0.76 | 0.83 | 0.83 | 0.77 |
| Anxiety and social distress | 0.70 | 0.77 | 0.77 | 0.82 | 0.82 | 0.74 |
| Overall accuracy | | | | | | |
| Total scale | 0.77 | 0.81 | 0.80 | 0.88 | 0.90 | 0.80 |
| Symptom distress | 0.76 | 0.80 | 0.81 | 0.88 | 0.90 | 0.79 |
| Interpersonal relations | 0.73 | 0.72 | 0.71 | 0.85 | 0.86 | 0.72 |
| Social role | 0.69 | 0.65 | 0.67 | 0.82 | 0.81 | 0.70 |
| Anxiety and social distress | 0.69 | 0.76 | 0.74 | 0.82 | 0.82 | 0.74 |
| Area under the curve | | | | | | |
| Total scale | 0.75 | 0.81 | 0.82 | 0.88 | 0.90 | 0.80 |
| Symptom distress | 0.75 | 0.80 | 0.81 | 0.88 | 0.90 | 0.79 |
| Interpersonal relations | 0.71 | 0.72 | 0.75 | 0.82 | 0.85 | 0.72 |
| Social role | 0.69 | 0.71 | 0.71 | 0.80 | 0.79 | 0.72 |
| Anxiety and social distress | 0.68 | 0.76 | 0.76 | 0.82 | 0.82 | 0.74 |

Note: Classification accuracy analyses were conducted on the cross-validation sample.
*Mean values of five stratified samples, $n = 3063$.

Table 4. Cut-off scores, areas under the curve, sensitivity and specificity of the Outcome Questionnaire-45 total scale for discrimination between the treatment populations

| | Private practice | Outpatient primary care | Outpatient secondary care | Part-time inpatient | Full-time inpatient | |
|---|---|---|---|---|---|---|
| Private practice | | 0.55 | 0.59 | 0.69 | 0.78 | Overall accuracy |
| Outpatient primary care | 70.2 | | 0.54 | 0.66 | 0.69 | |
| Outpatient secondary care | 72.2 | 75.6 | | 0.60 | 0.63 | |
| Part-time inpatient | 78.7 | 81.8 | 84.5 | | 0.52 | |
| Full-time inpatient | 80.2 | 83.4 | 86.2 | 91.2 | | |
| | | | Cut-off scores | | | |
| Private practice | | 0.52 | 0.57 | 0.67 | 0.82 | Specificity |
| Outpatient primary care | 0.56 | | 0.53 | 0.65 | 0.69 | |
| Outpatient secondary care | 0.59 | 0.53 | | 0.60 | 0.63 | |
| Part-time inpatient | 0.66 | 0.62 | 0.57 | | 0.53 | |
| Full-time inpatient | 0.69 | 0.64 | 0.56 | 0.46 | | |
| | | | Sensitivity | | | |

Note: Classification accuracy analyses were conducted on the cross-validation sample.

These patients were also the youngest. The relatively young age might have to do with the fact that the facilities were treatment centres for personality disorders, and that personality disorder severity declines somewhat when patients become older (Lenzenweger, 1999), or younger people may have fewer obligations (e.g., work and family) that could complicate engaging in long-term intensive treatment. Outpatient secondary care encompasses for a large part patients that could not adequately be treated in the outpatient primary care group and consequently show higher OQ-45 scores than primary care outpatient groups. The patients who are treated in private practices form a very heterogeneous group. Many of them are relatively high functioning and seek treatment for different reasons than psychiatric symptoms (e.g., obtaining insight in their personality). Hence, this group has lower OQ-45 scores on average than the other therapy groups.

One of the purposes of this study was to obtain an overall cut-off score for the OQ-45 that was based on a larger variety in patient populations. The OQ-45 is in use in other patient groups than secondary care outpatients, and this should be reflected in the cut-off score. Including other patient populations did not raise the cut-off for normal functioning much (one point, from 55 to 56), but will result in slightly higher numbers of recovered patients. However, the difference with the US cut-off score (63) is still quite large. A possible explanation for the difference between the Dutch and US cut-off is that the US patient populations report more problems than the Dutch outpatients. This might be due to cultural differences in scoring, but also to differences in the healthcare system. Psychotherapy is not always covered by insurance companies in the USA and not everyone has health insurance, whereas health insurance is mandatory in the Netherlands, and does cover (at least partially) psychotherapy. New developments in health care, with the introduction of the Affordable Care Act in the USA and the increasing amount of co-pay in the Netherlands, may make it necessary to obtain new normative data in a few years. An alternative explanation might be that different types of patient samples were used in the US and Dutch studies. It would be challenging to determine to what degree samples from different countries are comparable, which makes it difficult to judge equality between language versions.

### Application of Cut-Off Scores

The sensitivity, specificity, overall accuracy and AUCs for distinction between the normal population and most patient populations were good, although they were slightly lower for the private practice group. This lower precision could be caused by a more heterogeneous private practice population. It can be concluded that it will generally be correct using these cut-off scores for discrimination from the normal population. The cut-off score between the normal and combined patient population can be satisfactorily used as an element in the decision-making process for ending psychotherapy.

Additionally, cut-off scores for the total OQ-45 scale between the several therapeutic populations are presented. These cut-off scores might be used for treatment allocation, as well as an indication for migration from a more intensive treatment to a lighter treatment or vice versa. It should be noted that for the distinction between the patient groups, many AUCs are too low (range 0.54–0.73) to justify stringent use of the cut-off scores for triage purposes. However, combined with additional information about the patients (e.g., previous episodes and comorbidity), it may aid clinical decision-making to use cut-offs between patient populations.

### Limitations

This study has a number of limitations. First, a small, significant difference between the two normal samples was observed. This difference may be caused by the exclusion of participants who reported to receive therapy, but the exclusion of 24 participants is insufficient to account for the total effect. Another reason for the difference may be the way of sampling. Sampling from non-profit and business organizations implies that the participants have work, and that may result in lower scores. Differences in demographic characteristics may account for the different OQ-45 scores across the two normative subsamples. However, it could also be argued that by applying multiple sampling methods, a more representative normal population sample has been achieved, as methods may differ in which subpopulations they can effectively target. Another issue with the normal sample is that it included more older participants than the patient populations. As older people tend to have slightly lower OQ-45 scores than younger people (Lambert *et al.*, 1996), this may result in slightly lower cut-off scores for normal functioning.

Second, we had no proper access to information on education for the outpatient samples and no access to information on marital status, work situation, ethnicity and diagnoses for all samples. Information on these aspects may be particularly relevant for the normative sample, to determine whether the sample characteristics are consistent with the intended normative population.

Third, data have been collected over a period of several years, in which many changes in mental health care have taken place. One of the consequences is that symptom severity in patient populations might have shifted over time. Additionally, since mental healthcare reform is still going on, norms and cut-off scores are likely to be of limited tenability when patient populations continue to shift.

Fourth, the overall cut-off score for normal functioning has been estimated based on the assumption that the proportion of primary and secondary care was equally divided across outpatients; and similarly for day patients and inpatients. However, these numbers might not be correct, and if more information on the proportion of patients becomes available, the overall cut-off score might have to be re-estimated.

Finally, although we found that OQ-45 scores increased when the intensity of treatment increased, it should be noted that it is not possible to determine if the level of care is an indicator for patient severity, and as such cannot be considered as validation of the OQ-45. Allocations to treatment could be influenced by a number of factors, including if one would be able to partake in long-term day patient or inpatient therapy.

### *Conclusion*

In this study, the reported cut-off scores of the OQ-45 have shown to differentiate satisfactorily between the normal population and a broad range of treatments. The cut-off scores can be used to support decisions regarding the continuation of treatment. The new overall cut-off score between the clinical and normal population was calculated to be 56. For this cut-off score, we applied the results of an overall patient population with a balanced representation of the several patient populations. The new reliable change index is 18, also based on this balanced patient population.

## ACKNOWLEDGEMENTS

We would like to thank mental health care providers Indigo Castricum, Amici, GGZ Dijk & Duin, GGZ Noord-Holland-Noord, De Viersprong and PsyQ The Hague, for providing data on the OQ-45.

## REFERENCES

Amble, I., Gude, T., Stubdal, S., Oktedalen, T., Skjorten, A. M., Andersen, B. J., … Wampold, B. E. (2014). Psychometric properties of the Outcome Questionnaire-45.2: The Norwegian version in an international context. *Psychotherapy Research*, *24*(4), 504–513. doi: 10.1080/10503307.2013.849016

Bhugra, D., Gupta, S., Bhui, K., Craig, T., Dogra, N., Ingleby, J. D., … Tribe, R. (2011). WPA guidance on mental health and mental health care in migrants. *World Psychiatry*, *10*(1), 2–10.

Chiappelli, M., Lo Coco, G., Gullo, S., Bensi, L., & Prestano, C. (2008). The outcome questionnaire 45.2. Italian validation of an instrument for the assessment of psychological treatments. *Epidemiologia e Psichiatria Sociale*, *17*(2), 152–161.

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

De Jong, K., Nugter, M. A., Lambert, M. J., & Burlingame, G. M. (2009). *Handleiding voor de afname en scoring van de Outcome Questionnaire -45 (OQ-45)*. Salt Lake City, UT: OQ Measures LLC.

De Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E. A., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology & Psychotherapy*, *14*, 288–301. DOI:10.1002/cpp.529.

De Jong, K., Timman, R., Hakkaart-Van Roijen, L., Vermeulen, P., Kooiman, K., Passchier, J., & Busschbach, J. V. (2014). The effect of outcome monitoring feedback to clinicians and patients in short and long-term psychotherapy: A randomized controlled trial. *Psychotherapy Research*, *24*, 629–639. DOI:10.1080/10503307.2013.871079.

Digiuni, M., Jones, F. W., & Camic, P. M. (2013). Perceived social stigma and attitudes towards seeking therapy in training: a cross-national study. *Psychotherapy*, *50*(2), 213–223. DOI:10.1037/a0028784.

Durham, C. J., McGrath, L. D., Burlingame, G. M., Schaalje, G. B., Lambert, M. J., & Davies, D. R. (2002). The effects of repeated administrations on self-report and parent-report scales. *Journal of Psychoeducational Assessment*, *20*(3), 240–257. DOI:10.1177/073428290202000302.

Finch, A. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology and Psychotherapy*, *8*, 231–242.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic. *Radiology*, *143*(1), 29–36.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *JCCP*, *59*(1), 12–19. DOI:10.1037//0022-006X.59.1.12.

Lambert, M. J. (2005). Emerging methods for providing clinicians with timely feedback on treatment effectiveness: An introduction. *Journal of Clinical Psychology*, *61*(2), 141–144. DOI:10.1002/jclp.20106.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., … Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy*, *3*(4), 249–258. doi: 10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S.

Lambert, M. J., Whipple, J. L., Smart, D. W., D. A., V., 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), 49–67.

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

Lenzenweger, M. (1999). Stability and change in personality disorder features: The longitudinal study of personality disorders. *Archives of General Psychiatry*, *56*(11), 1009–1015. DOI:10.1001/archpsyc.56.11.1009.

Lo Coco, G., Chiappelli, M., Bensi, L., Gullo, S., Prestano, C., & Lambert, M. J. (2008). The factorial structure of the Outcome Questionnaire-45: A study with an Italian sample. *Clinical Psychology and Psychotherapy*, *15*(6), 418–423. DOI:10.1002/cpp.601.

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M. A., Jorasz, C., … Tschitsaz-Stucki, A. (2013). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, *23*, 14–24.

Moessner, M., Gallas, C., Haug, S., & Kordy, H. (2011). The Clinical Psychological Diagnostic System (KPD-38): Sensitivity to change and validity of a self-report instrument for outcome monitoring and quality assurance. *Clinical Psychology & Psychotherapy*, *18*, 331–338. DOI:10.1002/cpp.717.

Newnham, E. A., Hooke, G. R., & Page, A. C. (2010). Progress monitoring and feedback in psychiatric care reduces depressive symptoms. *Journal of Affective Disorders*, *127*(1–3), 139–146.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed. ). New York: McGraw-Hill.

Probst, T., Lambert, M. J., Loew, T. H., Dahlbender, R. W., Göllner, R., & Tritt, K. (2013). Feedback on patient progress and clinical support tools for therapists: Improved outcome for patients at risk of treatment failure in psychosomatic in-patient therapy under the conditions of routine practice. *Journal of Psychosomatic Research*, *75*, 255–261.

Probst, T., Lambert, M. J., Loew, T. H., Dahlbender, R. W., & Tritt, K. (in press). Extreme deviations from expected recovery curves and their associations with therapeutic alliance, social support, motivation, and life events in psychosomatic in-patient therapy. *Psychotherapy Research 25*(6), 714–723.

Qin, Y.-F., & Hu, S.-J. (2008). Usability report of Outcome Questionnaire 45.2 in part of Chinese sample. *Chinese Journal of Clinical Psychology*, *16*, 138–140.

Schauenburg, H., & Strack, M. (1999). Measuring psychotherapeutic change with the symptom checklist SCL 90 R. *Psychotherapy and Psychosomatics*, *68*(4), 199–206.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, *78*, 298–311.

Simon, W., Lambert, M. J., Harris, M. W., Busath, G., & Vazquez, A. (2012). Providing patient progress information and clinical support tools to therapists: Effects on patients at risk of treatment failure. *Psychotherapy Research*, *22*, 638–647.

Soeteman, D., Hakkaart-van Roijen, L., Verheul, R., & Busschbach, J. J. (2008). The economic burden of personality disorders in mental health care. *Journal of Clinical Psychiatry*, *69*(2), 259–265.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Thomson, S., Osborn, R., Squires, D., & Jun, E. (2013). International profiles of health-care systems. The Commonwealth Fund, November 2013. Retrieved from http://www.commonwealthfund.org/publications/fund-reports/2013/nov/international-profiles-of-health-care-systems

Tingey, R. C., Lambert, M. J., Burlingame, G. M., & Hansen, N. B. (1996). Assessing clinical significance: Proposed extensions to method. *Psychotherapy Research*, *6*(2), 109–123.

Wennberg, P., Philips, B., & De Jong, K. (2009). The Swedish version of the Outcome Questionnaire (OQ-45): Reliability and factor structure in a substance abuse sample. *Psychology and Psychotherapy*, *83*(Pt 3), 325–329. DOI:10.1348/147608309X478715.