The Symptom Questionnaire-48 (SQ-48) as an outcome measure for psychological distress in psychiatric outpatients: test-retest reliability and responsiveness to therapeutic change

Authors: Ingrid Carlier, Viktória Kovács, Erik Giltay, Bert van Hemert

Date: May 2014
Title: The Symptom Questionnaire-48 (SQ-48) as an outcome measure for psychological distress in psychiatric outpatients: test-retest reliability and responsiveness to therapeutic change

Status / citation: Internal Report LUMC (2014)

Authors: Ingrid Carlier, Viktória Kovács, Erik Giltay, Bert van Hemert

Date / place: May 2014, Leiden

Contact details: Leiden University Medical Centre (LUMC), Department of Psychiatry, Albinusdreef 2, Postzone Bi-P, 2300 RC Leiden, The Netherlands. Tel.: 0031(0)71-5265237, fax: 0031(0)71-5248156, E-mail address: I.V.E.Carlier@lumc.nl (I. Carlier).
PREFACE

The research described in this internal report is conducted under the supervision of the Department of Psychiatry of the Leiden University Medical Centre (LUMC). It is a follow-up study to the psychometric properties of the SQ-48. The SQ-48 was developed in 2011 by the LUMC Department of Psychiatry (Head at that time: Prof. dr. Frans Zitman) in collaboration with mental health care provider Rivierduinen. The SQ-48 is meant as a public domain self-report questionnaire for the screening of psychological distress.

In a previous study, the SQ-48 was validated in both clinical and nonclinical samples, and demonstrated good internal consistency as well as good convergent and divergent validity (Carlier et al., 2012*). Further research with the SQ-48 on the test-retest reliability and the responsiveness to change was recommended, which is therefore the focus of the present research described in this report (publication in preparation).

We very gratefully acknowledge the essential contribution made by the participants of the present study, as well as the following participating mental health care providers of the partnership SynQuest (in alphabetical order): Breburg, Delfland, Dimence, Noord-Holland-Noord, inGeest, and Rivierduinen. We also thank Margot Metz of Breburg who coordinated the multicentre data collection regarding the responsiveness of the SQ-48. Finally, we thank Marian Lucas and Marieke de Wit of the Department of Psychiatry of the LUMC for their coordination of the data collection regarding the test-retest reliability of the SQ-48.

SUMMARY

The assessment of psychological distress is important, because it may help to monitor treatment effects and predict treatment outcomes. Since no psychological distress instrument was available that also measures work functioning and positive psychological constructs like vitality, we previously developed the 48-item Symptom Questionnaire (SQ-48). The SQ-48 is meant as a public domain self-report questionnaire for the screening of psychological distress. We demonstrated that the SQ-48 has good internal consistency as well as good convergent and divergent validity among both clinical and nonclinical samples (Carlier et al., 2012a). The present study among psychiatric outpatients in a routine clinical setting, describes additional psychometric properties of the SQ-48. A test-retest design was used to examine its test-retest reliability (N=43 paired assessments). A longitudinal design was used to compare its responsiveness to change before and after a certain treatment period: comparison of the SQ-48 with the Brief Symptom Inventory (BSI, N=97) and comparison of the SQ-48 with the Outcome Questionnaire-45 (OQ-45, N=109). The results showed that the SQ-48 has excellent test-retest reliability and good responsiveness to therapeutic change. There were no significant differences between the SQ-48 and either the OQ-45 and the BSI in terms of responsiveness. In sum, the SQ-48 is a psychometrically sound self-report instrument to be used as a public domain screening or monitoring tool, as a benchmark tool, and for research purposes.
1. INTRODUCTION

Self-report instruments of psychological distress are widely used as psychiatric screening tools in clinical and research settings. The main purpose of such instruments is the assessment of the aggregate level of nonspecific psychological distress (Dohrenwend et al., 1980; Oakley Browne et al., 2010). In the literature, multiple related concepts are used interchangeably to refer to what these instruments measure: psychological-, psychosocial-, emotional-, affective-, mental-, global-, symptom-, psychiatric distress, or psychopathology. The assessment of psychological distress is important in (mental) health care, because of its relevance for the planning of treatment and prediction of treatment outcomes (Ritsner et al., 2002).

In addition to the measurement of psychological distress, attention should also be paid to work functioning and more positive constructs such as vitality or optimism. The importance of these constructs has already been demonstrated in various ways (Giltay et al., 2004; Giltay et al., 2006; Kubzansky and Thurston, 2007; Burdick et al., 2010). In most cases, psychopathology significantly impairs work functioning and adequate treatment may substantially reduce work impairment (Baune et al., 2007). Since no psychological distress instrument was available that also measures work functioning and vitality, we previously developed such an instrument: the 48-item Symptom Questionnaire (SQ-48). The SQ-48 has been validated among both clinical and nonclinical samples, and has good internal consistency as well as good convergent and divergent validity (Carlier et al., 2012a). The present research with the SQ-48 focusses on the test-retest reliability and the responsiveness to change.

Methods to evaluate the responsiveness of questionnaires usually involve the administration of the questionnaire before and after a certain time (treatment) period during which it is expected that symptomatology or functioning will improve (Davidson & Keating, 2002). However, to be of value, a questionnaire should also be stable when no clinically important change occurs (Beaton et al., 1997). This reproducibility can be ascertained using test-retest reliability, which concerns the degree to which repeated measurements in stable persons (no treatment started yet) provide similar results (Cassard et al., 1995). For test-retest reliability, a time-interval is needed that is sufficiently short (e.g. one or two weeks) to support the assumption that the condition remains stable, and sufficiently long to prevent a strong effect of recall (Spies-Dorgelo et al., 2006).

In general, research on the responsiveness to therapeutic change of self-report psychological distress instruments is very scarce, especially in the context of Routine Outcome Monitoring (ROM) (McClendon et al., 2011). One of the exceptions is the study of Moessner et al. (2011), who examined the suitability of the KPD-38 ("Klinisch Psychologisches Diagnosesystem 38") to detect therapeutic change in comparison with two widely used outcome measures: the OQ-45 (Outcome Questionnaire
and the BSI (Brief Symptom Inventory; Derogatis, 2000). Both OQ-45 and BSI showed higher sensitivity to change than the KPD-38 (Moessner et al., 2011).

The present research reports on the test-retest reliability and responsiveness to therapeutic change of the SQ-48 within a diversity of psychiatric outpatient populations. As in the study of Moessner et al. (2011), the responsiveness of the SQ-48 was compared with the OQ-45 and with the BSI.
2. METHODS

2.1. Design and procedure

The present study design was naturalistic and observational within routine clinical setting, using a web-based ROM programme. The research was conducted by the Department of Psychiatry of the Leiden University Medical Centre (LUMC).

More specifically, a test-retest design was used to examine test-retest reliability, where psychiatric outpatients were tested twice with the SQ-48, 1 week apart. The data were collected at the outpatient clinic of the Department of Psychiatry of the LUMC.

In addition, a longitudinal design was used to examine responsiveness to change. Before and after a certain treatment-period, psychiatric outpatients were given the SQ-48 and a second comparable questionnaire (BSI or OQ-45, see below). With regard to the content of treatment, evidence-based treatment was provided to all patients. Treatment interventions were not controlled or influenced by the research team. Responsiveness data were collected at six Dutch regional mental health care institutions (Breburg, Delfland, Dimence, Noord-Holland-Noord, inGeest, Rivierduinen), all participating in the so called SynQuest partnership.

The Medical Research Ethics Committee of the LUMC approved the general study protocol including ROM. This comprehensive protocol (PAREL: Psychiatric Academic Registration Leiden database) safeguards anonymity of the ROM-participants and ensures proper handling of the data. The protocol is available on request for patients. Since ROM is set up as part of the treatment, informed consent is not required for them. In case patients object to the use of their anonymous data for research purposes, these data will be removed from the research database. Additionally, the necessary approval for the present study was obtained from the Ethics Committee of the participating mental health care institutions of SynQuest.

2.2. Participants

The study population consisted of three separate cohorts, which had to do with the specific goals of the present study. The three cohorts are described below. For further details: see also Table 1 of the results.
Cohort 1: Comparing the responsiveness of the SQ-48 and the OQ-45

For this comparison, the sample consisted of 109 psychiatric outpatients (age range 19 - 63 years) from three Dutch mental health care institutions (SynQuest partnership, in alphabetic order): Breburg, Dimence, and Noord-Holland-Noord. The patients were treated between December 2012 and November 2013 for various psychiatric disorders (see Table 1). All patients filled in the questionnaires SQ-48 and OQ-45 twice: before treatment and about 19 weeks later. During each measurement, both questionnaires were presented. For further details regarding the questionnaires: see Table 2 of the results and below.

Cohort 2: Comparing the responsiveness of the SQ-48 and the BSI

For this comparison, the sample consisted of 97 psychiatric outpatients (age range 19 - 71 years) from three other Dutch mental health care institutions (SynQuest partnership, in alphabetic order): Delfland, GGZ inGeest, and Rivierduinen. The patients were treated between February 2013 and November 2013 for various psychiatric disorders. Patients filled in the questionnaires SQ-48 and BSI twice: before treatment and about 19 weeks later. During each measurement, both questionnaires were presented. For further details regarding the questionnaires: see Table 2 of the results and below.

Cohort 3: Test-retest reliability of the SQ-48

This sample consisted of 43 psychiatric outpatients (age range 20 - 83 years), referred to the Department of Psychiatry of the LUMC in 2013. The SQ-48 was completed twice by the patients before the start of their treatment, with a week between the two consecutive measurements.

2.3. Routine Outcome Monitoring (ROM)

In general, data were collected during a 1 to 2 hours ROM baseline / follow-up assessment at the mental health care centres. The ROM assessment could consist of a face-to-face psychiatric interview (optional, Mini-International Neuropsychiatric Interview or MINI; Sheehan et al., 1998), the administration of observer-rated instruments (optional), and self-report questionnaires (generic and disorder-specific). The diagnosis was determined by the clinical diagnosis of a psychiatrist / psychotherapist (or by MINI). For more details about the web-based ROM programme, we refer to
relevant publications (de Beurs et al., 2011; van Noorden et al., 2010; Carlier et al., 2012b; see also www.lumc.nl/psychiatry/ROM-instruments). For the purpose of the present study, we used the data of the following generic questionnaires: SQ-48, OQ-45, BSI. The procedure for the three self-report instruments was the same in all participating institutions. We chose the BSI and OQ-45 (to compare with SQ-48), because they focus on the same construct as the SQ-48 (psychological distress) and they are often used as generic outcome measures in ROM.

2.4. Outcome measures

2.4.1. Symptom Questionnaire-48 (SQ-48)

The SQ-48 was intended as a generic screening questionnaire that assesses common psychopathological symptoms useful for diagnostic screening. It also measures work functioning and vitality. More specifically, there were nine subscales in the SQ-48. Five subscales covered aspects of psychopathology: Depression ("MOOD" subscale: items 3, 7, 13, 19, 38, 40), Anxiety, ("ANXI" subscale: items 24, 28, 33, 41, 46, 48), Somatization/Somatic complaints ("SOMA" subscale: items 1, 5, 11, 17, 22, 26, 31), Social Phobia ("SOPH" subscale: items 23, 27, 32, 36, 45), and Agoraphobia ("AGOR" subscale: items 4, 8, 14, 25). In addition, four subscales were constructed to assess specific aspects of behaviour and/or functioning: Aggression ("AGGR" subscale: items 10, 16, 21, 43), Cognitive problems/complaints ("COGN" subscale: items 2, 6, 39, 44, 47), Work ("WORK" subscale: items 9, 15, 20, 30, 35), and Vitality ("VITA" subscale: items 12, 18, 29, 34, 37, 42). Each item is rated by the respondent on a 5-points Likert-scale (0: ‘Never’, 1: ‘Rarely’, 2 ‘Sometimes’, 3: ‘Often’, 4: ‘Very often’). Mean administration time was 5.4 minutes (SD=1.4).

The scores from the subscales are summed to create a total score of the SQ-48 that represents one’s overall level of psychological distress. The total score is the sum of 37 items (excluding WORK and VITA subscales), thus composed of 7 of the 9 subscales, and ranging from 0 to 148, with higher scores indicating more psychological distress. The subscales WORK and VITA were not included in the present analysis, because of their poor construct overlap with the other subscales and their different scoring (WORK: answering option “not applicable”; VITA: reverse scoring of items).

Respondents received the following instruction: “Try to answer the following statements honestly and accurately. Please indicate what applies best to you. There are no ‘right’ or ‘wrong’ answers. Give the answer that best expresses how often you have felt that way in the last week, including today. The answer which comes to your mind first, is often the best answer. Note: If you did not work or study or have not been able to do so, then you can skip questions 9, 15, 20, 30 and 35".
The scoring of the SQ-48 items is as follows. For the score of all subscales, the scores of the relevant items must be added. The 48 items are scored 0-4.

The Dutch SQ-48 was translated into English, according to guidelines for translation and cultural adaptation of questionnaires (Wild et al., 2005). Both English and Dutch SQ-48 are available as Supplementary material associated with the first article on the SQ-48, and they can be found in the online version (see Carlier et al., 2012a).

2.4.2. Outcome Questionnaire-45 (OQ-45)

The OQ-45 is a 45-item self-report measure developed specifically for the purpose of tracking and assessing client outcomes in a therapeutic setting (Lambert et al., 1996, Lambert & Hawkins, 2004; de Beurs et al., 2005; Lo Coco et al., 2008; Moessner et al., 2011). It was designed to assess three main areas of functioning that are considered important in defining outcome: symptomatic functioning (mainly anxiety and depression), interpersonal problems (friendship and family relations) and social role performance (work adjustment and quality of life) (Lo Coco et al., 2008). In keeping with this, the OQ-45 has three subscales: the Symptom Distress subscale (25 items), the Interpersonal Relations subscale (11 items), and the Social Role subscale (9 items). The OQ-45 is scored using a five-point Likert scale (0 = never, 1 = rarely, 2 = sometimes, 3 = frequently, 4 = almost always). For the present analysis, we only used the Symptom Distress subscale (OQ-SD), which yields a possible range of scores from 0 to 100, with higher scores indicating more psychological distress. Psychometric properties of the OQ-45 appear to be strong across a variety of cultures (Lo Coco et al., 2008).

2.4.3. Brief Symptom Questionnaire (BSI)

The Brief Symptom Inventory (BSI), a short version of the Symptom Checklist-90 (SCL-90), is a validated self-report questionnaire with 53 items. These items define a broad spectrum of psychological symptoms in the preceding 7 days (Derogatis & Melisaratos, 1983; Derogatis, 2000; de Beurs & Zitman, 2006). The BSI consists of 9 subscales: Somatization (SOM), Obsessive-Compulsive (O-C), Interpersonal Sensitivity (I-S), Depression (DEP), Anxiety (ANX), Hostility (HOS), Phobic Anxiety (PHOB), Paranoid Ideation (PAR), and Psychoticism (PSY). Item scores range from 0 (“not-at-all”) to 4 (“extremely”). The total scores is calculated as an average of all 53 items, with higher scores indicating more severe psychopathology or psychological distress.
2.5. Statistical analysis

Baseline categorical characteristics are presented as number (percentage), continuous variables are presented as mean (± standard deviation SD, with range).

Internal consistency was determined with Cronbach’s alpha coefficient. In order to demonstrate homogeneity of items, Cronbach’s alpha coefficient should have a value between 0.70 and 0.90.

Test-retest reliability (cohort 3) was determined using Intraclass Correlational Coefficients (ICCs, model “random”, type “consistency”). Interpretation of ICCs: 0.50<r<0.70 is medium, 0.70<r<0.90 is high, 0.90<r<1.00 is very high correlation.

Correlations were calculated at baseline and follow-up between SQ-48 and OQ-45, and between SQ-48 and BSI. To calculate the Residual Change Score (RCS) a regression analysis was performed. Baseline was the independent variable and follow-up was the dependent variable. Subsequently the correlation between the two Residual Change Scores was calculated.

The most efficient set of statistic measuring responsiveness is still a matter of debate (Davidson & Keating, 2002; Angst et al., 2008). Therefore, researchers often use several (two or three) comparable sets of statistics. In the present study, we used the following three statistics: (1) ES for the effect size index of Cohen; (2) SRM, Standardised Response Mean; and (3) ΔT, pre- post difference of normalized T-scores. In order to calculate the ES, we divided the difference between the pre- and post-test by the SD of the pre-test score (Cohen, 1988). In order to calculate the SRM, we divided the difference between the pre- and post-test by the SD of the difference score (Streiner & Norman, 2003). Positive values reflect (standardized) improvements in the number of standard deviations of the baseline scores (ES) or the score differences (SRM) (Angst et al., 2008). Since the SRM and ES are conceptually largely similar, they are usually interpreted considering the same benchmarks (0.2 or less: small, 0.5: moderate, 0.8 or greater: large; Streiner & Norman, 2003; Luiz & Almeida, 2012). Responsiveness was calculated with the use of: BSI total score, OQ-45 subscale Symptom Distress (OQ-SD), SQ-48 total score (7 subscales, excluding WORK and VITA).

To compare the results of the various measurements, we used the (normalized) T-score. To calculate the normalized T-scores, the scores of baseline and follow-up measurement should be standardized. Then, the raw scores, which are found to be not normally distributed, must be transformed to yield a normal distribution. For converting raw scores to (normalized) T-scores, the conversion factors for both OQ-45 and BSI are available (de Beurs et al., 2005; de Beurs & Zitman, 2006; de Beurs et al., 2012). As the T-score for SQ-48 has not previously been calculated and the raw scores of SQ-48 were normally distributed, we used a formula to convert these raw scores into a linear T-score. Then we calculated the (normalized) T-scores for both baseline and follow-up.
measurement. The $\Delta T$, the difference between the pre-test and post-test T-scores, was calculated (Cohen, 1988). In addition, we conducted analysis of variance to detect any differences between the two instruments in sensitivity to change.

Finally, we compared the results of the questionnaires. We calculated how many patients improved compared to the baseline measurement (RC: Reliable Change). Also, we looked at the number of patients who recovered with respect to the baseline measurement (CSC: Clinical Significant Change). Both RC and CSC were computed according to Jacobsen et al. (1999). McNemar chi-squared test was used to demonstrate how the results, improved or recovered patients, differed between the questionnaires.

All analyses were performed using SPSS 20.0 for Mac (SPSS Inc., Chicago, Illinois), with a significance level of p<.05.
3. RESULTS

3.1. Demographics and baseline characteristics of the sample

The baseline characteristics of the patients in the three cohorts are shown in Table 1.

SEE TABLE 1

More women can be found in cohort 1 (62.4 %) and 2 (69.1 %), whereas cohort 3 was more balanced for gender (51.2%). The average age was around 40 years (cohorts 1 and 2) and 50 years (cohort 3). The interval between the measurements within cohort 1 and cohort 2 was almost similar (19.2 ± 7.9 vs. 18.6 ± 7.3 weeks). Because cohort 3 was used to calculate test-retest reliability, the interval between the two measurements was much shorter (7.2 ± 1.0 days). In cohort 1, most patients had anxiety disorders (39.4 %) or depressive disorders (34.9%), whereas in cohort 2 most patients had depressive disorders (42.3%). In cohort 3 (test-retest), most patients had other disorders (53.4 %), especially somatoform disorders.

3.2. Characteristics of the questionnaires

In Table 2, the characteristics of the questionnaires are described. In addition, the mean and standard deviation of the baseline measurement of patients are presented. We used the RCI and CO scores of BSI and OQ-45 from the appropriate manuals. For the total scale of the SQ-48, we calculated the RCI and CO on the basis of formulas from Jacobsen et al. (1999).

SEE TABLE 2

Table 3 shows the correlations between the raw scores on the questionnaires.

SEE TABLE 3

From Table 3, it becomes clear that correlations between the questionnaires are high on both baseline and follow-up. These results indicate that the questionnaires measure the same underlying construct (i.e. psychological distress).
3.3. Test-retest reliability of the SQ-48

Table 4 shows the results of the test-retest reliability analyses.

SEE TABLE 4

ICCs were calculated for the subscales and total scale of SQ-48. With the exception of the Aggression subscale (ICC = 0.65), there was an acceptable medium to very high test-retest reliability for the SQ-48 subscales (ICCs ranging from 0.79 to 0.91) and a very high test-retest reliability for the SQ-48 total score (ICC = 0.93).

3.4. Responsiveness to therapeutic change of the SQ-48 compared with BSI and OQ-45

Table 5 shows the mean scores, standard deviation, and (normalized) T-scores. We also calculated the effect size (ES), standardized response mean (SRM) and ΔT. In addition, the results of the analysis of variance are presented in the table.

SEE TABLE 5

With the analysis of variance, we tested if there was a difference between the mean T-scores. There was no statistically significant interaction between time and questionnaire. Therefore, there were no significant differences between the SQ-48 and either the OQ-SD and BSI in terms of responsiveness.

Finally, in Table 6, we see the amount of patients who improved and recovered compared to the baseline measurement.

SEE TABLE 6

With the McNemar chi-squared test, we tested the difference between the questionnaires. Both the number of improved patients and the number of recovered patients were comparable, without any significant differences between the SQ-48 and either the OQ-SD and BSI.
4. CONCLUSIONS AND DISCUSSION

In a previous study on the development and validation of the SQ-48, we demonstrated that it is a psychometrically sound self-report measure for psychological distress, work functioning, and vitality (Carlier et al., 2012a). The present study showed that the SQ-48 has excellent test-retest reliability and a good responsiveness to therapeutic change. In sum, the combined results of these two studies justify the use of the SQ-48 as a public domain screening or monitoring tool, as a benchmark tool, and for research purposes.

This study contributes to the additional psychometric properties of the SQ-48 as a generic instrument that, if desired, can be used as an outcome measure in the context of ROM. Routine monitoring of patient outcomes enables adjusting treatment procedures when suitable progress is not observed, or when adverse effects of treatment are detected (Lilienfeld, 2007; McClendon et al., 2011). Therefore, investigating how well an instrument captures therapeutic change is crucial. To detect change over time, there are two important conditions for an instrument (Beaton et al., 1997; Martin & Philippon, 2008). First, an instrument must be reliable (consistent responses when no change has occurred or the change was not expected), which is reflected in its test-retest reliability. Second, an instrument must be responsive to clinical change (when change has occurred and was expected because of treatment). Our results showed that the SQ-48 meets both required conditions to detect change over time in patients.

Concerning the test-retest reliability of an instrument, the choice of the time period between test and retest is crucial. For the present study, we chose for a period of one week between test and retest, which is consistent with the literature (Spies-Dorgelo et al., 2006; Strand et al., 2008). There is some discussion regarding whether test-retest reliability measures the psychometric properties of a questionnaire or the relative stability of the disease itself (DeVellis, 2003). For many diseases, stability cannot be guaranteed. This could also be the case for psychopathology, which is sometimes characterized with a considerable variability in the transient emotional states of patients. Despite this, values of correlation coefficients between 0.7 and 0.9 are generally considered to indicate satisfactory to good test-retest reliability (DeVellis, 2003). Our test-retest results of the SQ-48 fall within this range.

In order to determine the responsiveness of the SQ-48, we compared it with two relevant generic instruments: the BSI and the OQ-45. Comparison of the responsiveness of two scales only makes sense, if they measure more or less the same content and construct within the same domain (Angst et al., 2008). This means that the two scales should have a high construct overlap, which is most often quantified by a correlation between the two scales (Angst et al., 2008). A correlation of more than 0.5 is considered to be appropriate when selecting an external criterion for assessing responsiveness (Spies-Dorgelo et al., 2006). De Beurs et al. (2012) found a high correlation between the BSI (total scale) and the OQ-45 (Symptomatic Distress scale), implying that both questionnaires
pertain to related concepts such as psychological distress, psychopathology. Our previous study on the SQ-48 already showed a high correlation between the SQ-48 and the BSI (Carlier et al., 2012a). The present study also showed high residual change correlations between the SQ-48 with either the OQ-SD and BSI (r=0.82 for both). In addition, de Beurs et al. (2012) found that the BSI and the OQ-45 both had good responsiveness values. These values confirm previous indications that the OQ-45 and BSI are sensitive to change in symptoms (Lambert et al., 1996; Peterson & Michael, 2007; de Jong et al., 2008; Moessner et al., 2011). In comparing the responsiveness of the SQ-48 with both the BSI and OQ-45, the present study showed that the questionnaires were equally responsive.

In identifying change with a generic questionnaire such as the SQ-48, it is recommended that also disorder-specific measures are used. For instance, in case of depression, the Beck Depression Inventory (BDI-II; Beck et al., 1996) could also be used. It has been shown that disorder-specific instruments are often more responsive that generic ones, although this is not necessarily true for every specific-generic comparison (Ruta et al., 1998; Wiebe et al., 2003; Whynes, 2009). In general, a clinician will advisedly use a range of patient reported outcome indicators to provide a picture of overall change, and the SQ-48 may be a prudent addition to the repertoire of assessment scales. Notwithstanding a careful choice of scale, there will always be some individuals who do not have a sufficiently high initial score to enable change to be reliably detected over time (Davidson & Keating, 2002).

Some limitations of our study need to be taken into account when interpreting the results. No detailed information was available as to the specific treatments of the disorders in this study. Moreover, it is not possible to differentiate the effectiveness of different therapies, because of the lack of a control group (Angst et al., 2008; McClendon et al., 2011). So, it is not clear whether the resulting effect differences in the present study reflect changes in mental health due to the specific treatment intervention or due to the natural course of the disorder. Also, our test-retest sample size (cohort 3) is relatively small, though comparable to most studies regarding test-retest reliability (Sauder et al., 2014). The sample size for responsiveness of the SQ-48 (cohort 1 and 2) complies with the required criterion of 100 to 200 patients (Cohen, 1988), but had a relatively broad range for the time interval between both measurements.

Recommendations for future research with the SQ-48 are the following. Now that it has been demonstrated that the SQ-48 is sufficiently responsive to therapeutic change, further research could focus on the ability of it to predict the future course of symptomatology and functioning, and whether a shortened version of the SQ-48 retains its good psychometric properties. Other suggestions include research with the SQ-48 among psychiatric inpatients and psychiatric patients from other cultures. With regard to the latter, it could be examined whether there may be ethnic differences in responsiveness to therapeutic change (Bjorndson et al, 2003). In this context, it can be mentioned that we currently run validation studies with the SQ-48 translated into Turkish and Moroccan among both clinical and nonclinical samples (publication in preparation). In addition, there is currently also a
validation study in progress in Portugal with the Portuguese version of the SQ-48 (Ana Varela, personal communication).
REFERENCES


Table 1. Baseline characteristics of psychiatric outpatients in 3 cohorts

<table>
<thead>
<tr>
<th></th>
<th>Cohort 1 (n=109)</th>
<th>Cohort 2 (n=97)</th>
<th>Cohort 3 (n=43)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female gender (%)</td>
<td>62.4</td>
<td>69.1</td>
<td>51.2</td>
</tr>
<tr>
<td>Age (yr)</td>
<td>38.6±11.8 (19 - 63)</td>
<td>38.1±12.4 (19 - 71)</td>
<td>47.9±15.5 (20 - 83)</td>
</tr>
<tr>
<td>Interval between measurements</td>
<td>19.2 ± 7.9 (3 - 42) wk</td>
<td>18.6 ± 7.3 (4 - 37) wk</td>
<td>7.2 ± 1.0 (2 - 9) ds</td>
</tr>
<tr>
<td>Ethnicity (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Dutch</td>
<td>89.0</td>
<td>62.9</td>
<td>93</td>
</tr>
<tr>
<td>• Other</td>
<td>7.3</td>
<td>4.1</td>
<td>7</td>
</tr>
<tr>
<td>• Unknown</td>
<td>3.7</td>
<td>33.0</td>
<td>0</td>
</tr>
<tr>
<td>Psychopathology (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Anxiety disorders</td>
<td>39.4</td>
<td>33.0</td>
<td>14.0</td>
</tr>
<tr>
<td>• Unipolar depressive disorders</td>
<td>34.9</td>
<td>45.4</td>
<td>27.9</td>
</tr>
<tr>
<td>• Other</td>
<td>25.7</td>
<td>16.4</td>
<td>53.4</td>
</tr>
<tr>
<td>• No diagnosis</td>
<td>0.0</td>
<td>5.2</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Data are number (percentages) or mean ± standard deviation (SD), with range. Cohort 1 was used to test the responsiveness of SQ-48 and OQ-45 questionnaires, cohort 2 was used to test the responsiveness of the SQ-48 and BSI questionnaires, cohort 3 was used to test the test-retest reliability of the SQ-48 questionnaire.

Other psychopathology in cohort 1: somatoform disorders, personality disorders, disorders usually first diagnosed in infancy, childhood or adolescence, adjustment disorders, impulse-control disorders not elsewhere classified, dissociative disorders, mental disorders due to a general medical condition not elsewhere classified.

Other psychopathology in cohort 2: personality disorders, bipolar disorders, disorders usually first diagnosed in infancy, childhood or adolescence, impulse-control disorders not elsewhere classified, eating disorders.

Other psychopathology in cohort 3: somatoform disorders, substance-related disorders, bipolar disorders, personality disorders.
<table>
<thead>
<tr>
<th>(Sub)scale</th>
<th>#items</th>
<th>Score range</th>
<th>α</th>
<th>Mean</th>
<th>SD</th>
<th>RCI</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>37</td>
<td>0 – 148</td>
<td>0.90</td>
<td>61.80</td>
<td>19.63</td>
<td>14.4</td>
<td>42</td>
</tr>
<tr>
<td>SQ-48 (composed of 7 subscales*)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OQ-45 Symptom distress</td>
<td>25</td>
<td>0 - 100</td>
<td>0.88</td>
<td>53.0</td>
<td>15.5</td>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>BSI Total score</td>
<td>53</td>
<td>0 - 4</td>
<td>0.95</td>
<td>1.39</td>
<td>0.65</td>
<td>0.35</td>
<td>0.68</td>
</tr>
</tbody>
</table>

SQ-48: Symptom Questionnaire 48, OQ-45: Outcome Questionnaire 45, BSI: Brief Symptom Inventory
Mean: Mean score of patients before the start of treatment, SD: Standard deviation, RCI: Reliable Change Index
CO: Cut-Off score using the 5th percentiles from the NormQuest sample for the SQ-48 (Carlier et al., 2012), for BSI (de Beurs et al., 2012; Schulte-van Maaren et al., 2012) and for the OQ-45 (de Jong et al., 2007)
α: Cronbach’s Alpha
*: SQ-48 total score composed of the following 7 subscales: Mood, Anxiety, Somatization, Social Phobia, Agoraphobia, Aggression, and Cognitive problems. The subscales Work and Vitality were excluded, because of their poor construct overlap with the other subscales and their different scoring (Work: answering option “not applicable”; Vitality: reverse scoring of items)
### Table 3. Pearson's product-moment correlation coefficients between raw scores on the questionnaires

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Follow-up</th>
<th>Difference score</th>
<th>Residual Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQ-48 total &amp; OQ-SD</td>
<td>0.83</td>
<td>0.87</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>SQ-48 total &amp; BSI total</td>
<td>0.84</td>
<td>0.88</td>
<td>0.79</td>
<td>0.82</td>
</tr>
</tbody>
</table>

SQ-48: Symptom Questionnaire 48, OQ-SD: Outcome Questionnaire 45 subscale Symptom Distress, BSI: Brief Symptom Inventory total subscale

To calculate residual change score, a regression analysis was performed. Baseline was the independent variable and follow-up was the dependent variable. Subsequently the correlation between the two Residual Change Scores was calculated.
Table 4: Test-retest reliability of the SQ-48 (sub)scales using cohort 3 (N=43)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Items</th>
<th>Cronbach’s alpha at baseline</th>
<th>ICC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Aggression</td>
<td>4</td>
<td>0.76</td>
<td>0.65 (0.44 - 0.80)*</td>
</tr>
<tr>
<td>2. Agoraphobia</td>
<td>4</td>
<td>0.78</td>
<td>0.79 (0.65 - 0.88)*</td>
</tr>
<tr>
<td>3. Anxiety</td>
<td>6</td>
<td>0.92</td>
<td>0.91 (0.84 – 0.95)*</td>
</tr>
<tr>
<td>4. Cognitive complaints</td>
<td>5</td>
<td>0.79</td>
<td>0.90 (0.82 – 0.94)*</td>
</tr>
<tr>
<td>5. Mood</td>
<td>6</td>
<td>0.89</td>
<td>0.91 (0.84 – 0.95)*</td>
</tr>
<tr>
<td>6. Somatic complaints</td>
<td>7</td>
<td>0.84</td>
<td>0.87 (0.77 - 0.93)*</td>
</tr>
<tr>
<td>7. Social phobia</td>
<td>5</td>
<td>0.90</td>
<td>0.89 (0.80 - 0.94)*</td>
</tr>
<tr>
<td>SQ-48 total score(^a)</td>
<td>37</td>
<td>0.94</td>
<td>0.93 (0.87 – 0.96)*</td>
</tr>
</tbody>
</table>

ICC - Intraclass correlation  
*p <0.001  
\(^a\): SQ-48 total score composed of the following 7 subscales: Mood, Anxiety, Somatization, Social Phobia, Agoraphobia, Aggression, and Cognitive problems. The subscales Work and Vitality were excluded, because of their poor construct overlap with the other subscales and their different scoring (Work: answering option “not applicable”; Vitality: reverse scoring of items)
Table 5. Comparison of the results of repeated measurements with cohort 1 (SQ-48 total and OQ-SD) and cohort 2 (SQ-48 total and BSI total)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>Follow-up</th>
<th>F-test</th>
<th>SRM</th>
<th>ES</th>
<th>ΔT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>T x M</td>
<td></td>
</tr>
<tr>
<td>SQ-48 total &amp; OQ-SD</td>
<td>F-test (1, 108)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ-48 total (raw score)</td>
<td>71.0</td>
<td>22.3</td>
<td>62.2</td>
<td>27.3</td>
<td>4.143</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>SQ-48 total (T-score)</td>
<td>54.7</td>
<td>11.4</td>
<td>50.2</td>
<td>13.9</td>
<td>0.78</td>
<td>p = 0.38</td>
</tr>
<tr>
<td>OQ-SD (raw score)</td>
<td>53.0</td>
<td>15.5</td>
<td>46.8</td>
<td>17.2</td>
<td></td>
<td>0.52</td>
</tr>
<tr>
<td>OQ-SD (T-score)</td>
<td>53.1</td>
<td>9.8</td>
<td>49.2</td>
<td>10.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ-48 total &amp; BSI total</td>
<td>F-test (1, 96)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ-48 total (raw score)</td>
<td>61.8</td>
<td>19.6</td>
<td>51.7</td>
<td>22.9</td>
<td>20.550</td>
<td>p &lt;0.001</td>
</tr>
<tr>
<td>SQ-48 total (T-score)</td>
<td>50.0</td>
<td>10.0</td>
<td>44.8</td>
<td>11.7</td>
<td>0.423</td>
<td>p = 0.52</td>
</tr>
<tr>
<td>BSI total (raw score)</td>
<td>1.4</td>
<td>0.6</td>
<td>1.02</td>
<td>0.7</td>
<td></td>
<td>0.59</td>
</tr>
<tr>
<td>BSI total (T-score)</td>
<td>53.3</td>
<td>8.1</td>
<td>47.7</td>
<td>10.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

M: Mean score; SD: Standard deviation
SRM: Standardised Response Mean. In order to calculate the SRM, we divided the difference between the pre- and post-test by the SD of the difference score
ES: Effect Size. In order to calculate the ES, we divided the difference between the pre- and post-test by the SD of the pre-test score
ΔT: Difference in normalized T-score; T x M: Interaction effect between time and questionnaire
SQ-48: Symptom Questionnaire 48; OQ-SD: Outcome Questionnaire 45 subscale Symptom Distress; BSI: Brief Symptom Inventory
Table 6: Comparing Clinical Significance and Reliable Change Index between the questionnaires

<table>
<thead>
<tr>
<th></th>
<th>Reliable change</th>
<th>Clinical significant improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RCI</td>
<td>%</td>
</tr>
<tr>
<td><strong>SQ &amp; OQ-SD (cohort 1; n=109)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ-48 total</td>
<td>14.4</td>
<td>33.0</td>
</tr>
<tr>
<td>OQ-SD</td>
<td>10</td>
<td>24.8</td>
</tr>
<tr>
<td><strong>SQ &amp; BSI (cohort 2; n=97)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ-48 total</td>
<td>14.4</td>
<td>43.3</td>
</tr>
<tr>
<td>BSI total</td>
<td>0.35</td>
<td>46.4</td>
</tr>
</tbody>
</table>

*McNemar chi-squared test was used to demonstrate how the results, improved or cured patients, differ between the questionnaires.
RCI: Reliable Change Index
SQ-48: Symptom Questionnaire 48
OQ-SD: Outcome Questionnaire 45 subscale Symptom Distress
BSI: Brief Symptom Inventory