To Be or Not to Be Improved: Patients’ Perception of Symptom Improvement – Linking the SCL-90-R to Patient-Rated Global Improvement in a Large Real-World Treatment Sample

Falk Leichsenring a, c Ulrich Jaeger b Oliver Masuhr b Andreas Dally b Michael Dümpelmann b Christian Fricke-Neef b Carsten Spitzer b, c Christiane Steinert a, d

a Department of Psychosomatics and Psychotherapy, University of Giessen, Giessen, Germany; b Asklepios Clinic Tiefenbrunn, Tiefenbrunn, Germany; c Department of Psychosomatics and Psychotherapeutic Medicine, University of Rostock, Rostock, Germany; d International Psychoanalytic University, Berlin, Germany

Keywords
Clinimetrics · Improvement · Outcome · Patient perception · Pharmacotherapy · Psychotherapy

Abstract
Introduction: From both a clinical and research perspective, it is important to determine what constitutes a perceivable change in commonly used outcome measures. Objective: We aimed to do so for the Symptom Checklist-90-Revised (SCL-90-R). Methods: Patients from a large real-world sample treated with inpatient psychotherapy (n = 4,791) rated improvements in symptoms on a global 5-point Likert scale at discharge. These ratings were related to pre-post changes in the Global Severity Index (GSI) of the SCL-90-R by use of equipercentile linking. Results: A patient rating of 5 (“clearly improved”) was found to be equivalent to an absolute pre-post difference in the GSI of 0.67 or to a percentage improvement of 54%, with the latter corresponding to the common definition of response as a 50% reduction in symptoms. A rating of 1 (“clearly worse”) was equivalent to an increase in the GSI > 0.50 and to a percentage worsening > 55%. “Slightly improved” or “slightly worse” (ratings of 4 or 2) corresponded to pre-post changes in the GSI of 0.07 and 0.50. For severely disordered patients, larger changes were required for ratings of improvement, and for less severely disordered patients, larger changes were required for ratings of worsening. Results for depressive, anxiety, and personality disorders were widely consistent with those of the total sample. Conclusions: This study is the first to link patient ratings of improvement or worsening to changes in the SCL-90-R. Results are relevant to both the interpretation of changes in individual patients and of effect sizes in outcome research. Results require replication.

Introduction

What constitutes a perceivable change in symptoms is important from both a clinical and a research perspective. In research, estimates of a perceivable change or a minimal clinically important difference (MCID) can be used for planning randomized controlled trials and for inter-
interpreting their results [1–3]. Several studies empirically addressed the issue of a perceivable change by applying equipercentile linking analysis using an anchor-based approach [4–8]. For these analyses, global patient ratings can be used and related to either established self-report instruments or observer-rated instruments [6, 9]. Patient global ratings of change may differ from clinician ratings [1, 9, 10].

The Symptom Checklist-90-Revised (SCL-90-R) [11], a commonly used self-report instrument, has not yet been studied in this regard. For these reasons, we carried out a study relating global patient ratings of improvement or worsening to changes in the SCL-90-R in a large heterogeneous sample of real-world inpatients. Thus, this is one of the few studies linking two self-report ratings.

Materials and Methods

In the Asklepios Clinic Tiefenbrunn in Germany, patients are treated with psychodynamic therapy which may be temporarily complemented by pharmacotherapy in cases of severe depression or anxiety. Patients are routinely assessed by use of validated diagnostic instruments. In the included sample, ICD-10 diagnoses were assigned by trained raters to ensure reliability. General symptom distress was assessed by the Global Severity Index (GSI) of the SCL-90-R at admission and discharge [11]. The SCL-90-R is a frequently used 90-item self-report symptom inventory using a 5-point scale, corresponding to a GSI of 0–4 [11]. In addition, patients rated the global change in symptoms (global symptom change rating, GSCR) on a 5-point Likert scale ranging from 1 (clearly worse) over 2 (slightly worse), 3 (neither improved nor worse), and 4 (slightly improved) to 5 (clearly improved). The GSCR is routinely applied in many clinics in Germany [12]. In addition, the “Veränderungsfragebogen des Erlebens und Verhaltens” (VEV) was applied, a frequently used and validated self-report instrument [13] including 42 items asking patients to rate their present state (well-being) in comparison to a given point in time on a 7-point scale. The items are summed up to a total score, ranging from 42 (extreme worsening) to 294 (extreme improvement). In this study, patients rated their psychological state at discharge in comparison to admission.

For 4,791 patients, data of both the GSCR and the SCL-90-R at admission and discharge were available. Most patients suffered from depressive disorders (84%), anxiety disorders (45%), personality disorders (50%), or stress-related and adjustment disorders (37%) and were given more than one ICD-10 diagnosis (86%). All patients gave written informed consent. Anonymized data were used that collected as part of the routine assessment in the clinic of Tiefenbrunn.

For relating pre-post changes in the GSI to the GSCR, we used the equipercentile linking procedure to identify those scores on the GSI and GSCR that have the same percentile ranks. [14]. Using the percentile rank function of one scale and the inverse percentile rank function of the other, for every score of one scale, a score on the other scale that has the same percentile rank was identified.

Statistical analyses were performed with the SAS program EQUIPERCENTILE [15]. Analyses were carried out for the total sample and several diagnostic groups. For classifying patients into more or less severely disordered, the GSI at baseline was split at the median.

Results

In the total sample, the mean GSI at admission was 1.40 (SD 0.65), which was significantly reduced to 0.92 (SD 0.67) at discharge (t = 59.90, p < 0.0001). The mean pre-post change in the GSI was 0.48 (SD 0.56), corresponding to a pre-post effect size of d = 0.74 [16].

The Spearman rank correlation between the change in the pre-post GSI and GSCR at discharge was significant (n = 4,791, r = 0.43, p < 0.0001), with a significantly higher correlation (z = 7.21, p < 0.01) in the severely disordered patients (i.e., above the baseline GSI median of 1.37) than in the less severely disordered patients (r = 0.56, n = 2,370 vs. r = 0.40, n = 2,421). The Spearman rank correlation between the GSCR and the VEV was significant (r = 0.48, p < 0.0001), too, providing evidence for the validity of the GSCR. These correlations represent medium to large effect sizes [17], thus allowing for linking GSCR with change in the GSI.

According to the results, a rating of “clearly improved” (GSCR = 5), achieved by 34% of the total sample, corresponded to an absolute pre-post difference in the GSI of 0.67 (Table 1). Dividing this difference by the baseline SD results in a large within-group standardized mean difference (SMD) of 1.03 (0.67/0.65). A rating of “clearly worse” made by 3% (GSCR = 1), on the other hand, corresponded to a GSI pre-post change <–0.50 (SMD <–0.80). Differences between male (n = 1,773) and female (n = 3,018) patients were small (e.g., 0.62 vs. 0.70 for a rating of “clearly improved”). However, there were some clear differences between severely and less severely disordered patients (Table 1). For the former, a rating of “slightly improved” corresponded to a reduction in the GSI of 0.20 (SMD 0.43), whereas in the latter, the corresponding value was only 0.01 (SMD 0.03). The corresponding pre-post changes for perceiving a “clear improvement” were ≥1.02 compared to ≥0.43. For a perception of “clearly worse,” a larger increase in symptoms was required in the less severely disordered patients (SMD –1.88 vs. –1.00). For patients with depressive, anxiety, and personality disorders, results were similar to those of the total sample (Table 2), suggesting that diagnosis does not matter for perceived improvement or worsening.
### Table 1. Absolute GSI pre-post change corresponding to patient global rating of symptom change (GSCR): total sample and severely and less severely disordered patients

| GSCR   | 1    | ≥2   | ≥3   | ≥4   | 5    |
|--------|------|------|------|------|------|
| Total sample (n = 4,791) |      |      |      |      |      |
| Absolute GSI pre-post change | <−0.50 | ≥−0.50 | ≥−0.24 | ≥0.07 | ≥0.67 |
| Standardized mean pre-post effect size (SMD) | −0.80 | −0.80 | −0.37 | 0.11 | 1.04 |
| Relative GSI pre-post change, % | ≤−56 | ≥−55 | ≥−22 | ≥5 | ≥54 |
| Severely disordered patients (n = 2,370) |      |      |      |      |      |
| Absolute GSI pre-post change | <−0.44 | ≥−0.44 | ≥−0.12 | ≥0.20 | ≥1.02 |
| Standardized mean pre-post effect size (SMD) | −1.00 | −0.98 | 0.27 | 0.45 | 2.32 |
| Relative GSI pre-post change, % | ≤−24 | ≥−23 | ≥−7 | ≥10 | ≥56 |
| Less severely disordered patients (n = 2,421) |      |      |      |      |      |
| Absolute GSI pre-post change | <−0.62 | ≥−0.62 | ≥−0.33 | ≥0.01 | ≥0.43 |
| Standardized mean pre-post effect size (SMD) | −1.88 | −1.85 | −1.00 | 0.03 | 1.30 |
| Relative GSI pre-post change, % | ≤−23 | ≥−22 | ≥−7 | ≥10 | ≥56 |

GSCR: 1 = "clearly worse," ≥2 = "slightly worse," ≥3 = "neither improved nor worse," ≥4 = "slightly improved," 5 = "clearly improved."

### Table 2. Absolute and relative GSI pre-post change corresponding to patient global rating of symptom change (GSCR): depressive, anxiety, and personality disorders

| GSCR   | 1    | ≥2   | ≥3   | ≥4   | 5    |
|--------|------|------|------|------|------|
| Depressive disorders (n = 4,005) |      |      |      |      |      |
| Absolute GSI pre-post change | ≤−0.52 | ≥−0.51 | ≥−0.25 | ≥0.08 | ≥0.69 |
| Effect size d | −0.81 | −0.80 | −0.39 | −0.13 | 1.08 |
| SCL-Depression Scale | ≤−0.69 | >−0.69 | ≥−0.30 | ≥0.15 | 1.10 |
| Standardized mean pre-post effect size (SMD) | −1.08 | −1.08 | −0.47 | 0.23 | 1.72 |
| Relative GSI pre-post change, % | ≤−53 | ≥−53 | ≥−21 | ≥5 | ≥54 |
| Anxiety disorders (n = 2,077) |      |      |      |      |      |
| Absolute GSI pre-post change | ≤−0.51 | ≥−0.51 | ≥−0.20 | ≥0.08 | ≥0.73 |
| Effect size d | −0.80 | −0.78 | −0.31 | 0.13 | 1.14 |
| SCL-Anxiety Scale | ≤−0.80 | >−0.80 | ≥−0.50 | ≥−0.10 | ≥0.80 |
| Standardized mean pre-post effect size (SMD) | −1.25 | −1.25 | −0.78 | −0.16 | 1.25 |
| Relative GSI pre-post change, % | ≤−56 | ≥−55 | ≥−20 | ≥7 | ≥57 |
| Personality disorders (n = 2,238) |      |      |      |      |      |
| Absolute GSI pre-post change | ≤−0.51 | ≥−0.51 | ≥−0.25 | ≥0.08 | ≥0.73 |
| Standardized mean pre-post effect size (SMD) | −0.77 | −0.77 | −0.38 | 0.12 | 1.11 |
| Relative GSI pre-post change, % | ≤52 | ≥−51 | ≥−21 | ≥6 | ≥54 |

GSCR: 1 = "clearly worse," ≥2 = "slightly worse," ≥3 = “neither improved nor worse,” ≥4 = “slightly improved,” 5 = “clearly improved.”
Discussion

This study is the first to link patient ratings of improvement to changes in the SCL-90-R GSI [2, 4, 5, 8, 14, 18, 19]. Only minor differences emerged between depressive, anxiety, or personality disorders. These results suggest that patient-perceived improvement is a general phenomenon. However, differences were found between severely disordered and less severely disordered patients. For severely disordered patients, larger changes were required to perceive an improvement, and for less severely disordered patients, larger changes were required for a perception of worsening.

In the total sample relating the global patient rating of improvement (GCSR) to relative change in the GSI, a rating of “clearly improved” (GCSR = 5) was equivalent to a pre-post percentage change in the GSI of at least 54% (Table 1), corresponding to the usual definition of response by a 50% reduction. Thus, “clearly improved” may be regarded as indicating response. In studies using the Clinical Global Impression-Improvement (CGI-I) , a rating of 2 (“much improved”) corresponded to a reduction ≥50%, indicating response [6] and suggesting that a rating of “clearly improved” in the GCSR corresponds to “much improved” in the CGI-I.

For the HAM-D, an improvement by 7 points was found to be required for a change detectable by clinicians, corresponding to a CGI-I rating of “minimally improved” [6]. Moncrieff and Kirsch [19] transformed this difference in a pre-post effects size, reporting a large effect (0.875). In a recent study, Furukawa et al. [8] linked the HAM-D with the BDI and BDI-II. The authors found an improvement in the HAM-D of 7 points to be equivalent to an improvement in the BDI of 9 points and in the BDI-II of 11 points [8]. Using the standard deviations published by Furukawa et al. [8], 9 and 11 points in the BDI and BDI-II can be transformed into within-group effect sizes of 1.06 and 1.07, implying that not only for observer-rated measures such as the HAM-D but also for self-report instruments, large within-group effect sizes are required to be detectable by clinicians.

For the data presented here, it may be discussed whether a GCSR of “slightly improved” or “clearly improved” corresponds to a rating of “minimally improved” in the CGI-I. However, “clearly improved” seems to indicate response and to rather correspond to “much improved” in the CGI-I. Thus, a rating of “slightly improved” in the GCSR may be regarded as corresponding to “minimally improved” in the CGI-I. Even for severely disordered patients, only a medium pre-post effect size was required (0.45) for a perception of “slightly improved” (GCSR = 4), suggesting that patients detect smaller differences in improvement than clinicians [1]. In the study presented here, patient ratings were used for both the global rating of change and the change in the SCL-90-R.

Within-group changes have been suggested as thresholds for clinical significance and sample size calculation in randomized controlled trials [1, 3]. However, within-group changes are conceptually different from between-group differences [20]. It is an interesting and finally open question if and how pre-post changes, for example, in terms of within-group effect sizes, may be related to between-group effect sizes [7, 20, 21]. Acknowledging the difference between within-group and between-group differences, Sung and Gantz [21] argued that within-group effect sizes may be used as an estimate for between-group differences and may therefore be used for study planning [1]. The proportion of patients within each treatment group who experienced a change as large as or greater than the MCID (i.e., treatment responders) can be compared between treatments [3, 20]. In fact, there is considerable research addressing the relationship between within-person and between-person differences [1, 22]. Barrett et al. [1, p. 253] conclude from the results that there is a close correspondence between an MCID derived from between-person assessments and that derived from within-person self-assessments. If anything, within-group effect sizes can be expected to rather underestimate between-group differences. Thus, if an empirically found between-group effect size is not at least as large as the MCID, or more precisely, if the upper limit of its 95% CI is not at least as large as the MCID, the difference between the treatments is not clinically important [22]. If, on the other hand, the lower limit of the 95% CI of the between-group effect size is larger than the MCID, the effect is definitely clinically important [22].

Clinical importance, however, does not necessarily imply that the effect represents a worthwhile difference. For this designation, the between-group effect must be weighed against the side effects and costs of a specific treatment [1, 2]. The smallest worthwhile difference (SWD) represents the smallest amount of benefit that justifies the side effects and costs of a treatment [1, 2]. The SWD is by definition always equal to or greater than the MCID [1, p. 254]; otherwise, the effect is not clinically important. An inefficacious treatment, that is, a treatment whose between-group effect is smaller than the MCID compared to placebo, cannot be worthwhile.
If it is not superior to, for example, placebo, rather the placebo can be recommended for treatment, assuming that it is associated with lower costs and side effects. For antidepressants, this seems to be the case [19], but also for many forms of short-term psychotherapy, especially if effect sizes are adjusted for bias [23]. For a specific treatment, the SWD may be larger than the MCID if its side effects or costs are larger than those of an alternative treatment. If two treatments yield identical between-group effect sizes in comparison to placebo (or no difference in a head-to-head comparison with each other), it depends on their cost and side effects which of them can be recommended for treatment. The SWD may be smaller for the treatment associated with less costs or side effects. In a non-inferiority trial, this SWD may be used as a margin.

It is of note, that small standardized mean between-group differences (e.g., SMD = 0.30) in comparison to placebo (or an active treatment) may correspond to relevant differences in response rates (e.g., 20%). This is especially the case if vital events are involved (e.g., mortality). If no more efficacious treatments are available whose side effects and/or costs do not exceed those of the treatment in question, the treatment yielding a small difference in response rates (e.g., 20%) may be recommended. However, as noted above, this is only the case if its effect size in comparison to placebo is larger than the MCID.

Empirically derived changes are relevant to the interpretation of both changes in individual patients and of effect sizes in outcome research. They can serve as benchmarks against which treatment effects can be compared. In a next step, treatment effects need to be weighed against costs and risks to see whether treatments are worthwhile, consistent with a clinimetric approach [24].

Acknowledgement

We would like thank Dr. Johannes Hüsing (KKS Heidelberg) for statistical support. We would like to thank Dr. Toshiaki A. Furukawa for his helpful comments that moved the discussion considerably forward.

Statement of Ethics

The research was conducted ethically in accordance with the World Medical Association Declaration of Helsinki.

Conflict of Interest Statement

For all authors, there are no conflicts of interest to be declared.

Funding Sources

There are no funding sources to declare.

Author Contributions

F.L. and C.St. conceived the concept of the study and wrote the first version of the manuscript. U.J. and O.M. performed statistical analyses. All authors revised the manuscript for intellectual content.

References

1 Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. Med Decis Making. 2005 May-Jun;25(3):250–61.
2 Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol. 2012 Mar;65(3):253–61.
3 McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA. 2014 Oct;312(13):1342–3.
4 Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel R. Clinical implications of Brief Psychiatric Rating Scale scores. Br J Psychiatry. 2005 Oct;187(4):366–71.
5 Leucht S, Engel RR, Davis JM, Kissling W, Meyer Zur Capellen K, Schmauß M, et al. Equipercentile linking of the Brief Psychiatric Rating Scale and the Clinical Global Impression Scale in a catchment area. Eur Neuropsychopharmacol. 2012 Jul;22(7):501–5.
6 Leucht S, Fennema H, Engel R, Kaspers-Janssen M, Lepping P, Szegedi A. What does the HAMD mean? J Affect Disord. 2013 Jun;148(2-3):243–8.
7 Leucht S, Fennema H, Engel RR, Kaspers-Janssen M, Lepping P, Szegedi A. What does the MADRS mean? Equipercentile linking with the CGI using a company database of mirtazapine studies. J Affect Disord. 2017 Mar;210:287–93.
8 Furukawa TA, Reijnders M, Kishimoto S, Sakata M, DeRubeis RJ, Dimidjian S, et al. Translating the BDI and BDI-II into the HAMD and vice versa with equipercentile linking. Epidemiol Psychiatr Sci. 2019 Mar;29:e24.
9 Stefanovics EA, Rosenheck RA, Jones KM, Huang G, Krystal JH. Minimal Clinically Important Differences (MCID) in Assessing Outcomes of Post-Traumatic Stress Disorder. Psychiatr Q. 2018 Mar;89(1):141–55.
10 Hengartner MP, Pölderl M. Statistically Significant Antidepressant-Placebo Differences on Subjective Symptom-Rating Scales Do Not Prove That the Drugs Work; Effect Size and Method Bias Matter! Front Psychiatry. 2018 Oct;9:517.
11 Derogatis L. The SCL-90 Manual I: Scoring, administration and procedures for the SCL-90-R. Baltimore: Clinical Psychometric Research; 1977.

12 Heuft G, Senf W, Bell K, Cording C, Geyer M, Janssen P, et al. [Psy-Bado-Kernmodul einer Basisdokumentation in der Fachpsychotherapie]. Psychotherapeut. 1998;43(1):48–52.

13 Hinz A. Veränderungsfragebogen des Erlebens und Verhaltens – VEV. In: Geue K, Strauß B, Brähler E, editors. Diagnostische Verfahren in der Psychotherapie. Göttingen: Hogrefe; 2016. pp. 509–10.

14 Kolen MJ, Brennan RL. Test Equating: Methods and Practices 3ed. New York Springer; 2014.

15 Price LR, Lurie A, Wilkins C. EQUIPERCENT: A SAS program for calculating equivalent scores using the equipercentile method. Applied Psychological Measurement. 2001; 25(4):332.

16 Becker BJ. Synthesizing standardized mean-change measures. Br J Math Stat Psychol. 1988;41(2):257–78.

17 Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale: Lawrence Erlbaum; 1988.

18 Hermes ED, Sokoloff D, Stroup TS, Rosenheck RA. Minimum clinically important difference in the Positive and Negative Syndrome Scale with data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE). J Clin Psychiatry. 2012 Apr;73(4):526–32.

19 Moncrieff J, Kirsch I. Empirically derived criteria cast doubt on the clinical significance of antidepressant-placebo differences. Contemp Clin Trials. 2015 Jul;43:60–2.

20 Furukawa TA. Measuring clinical importance in a trial of interventions for mixed urinary incontinence. JAMA. 2020 Feb;323(5):479.

21 Sung VW, Gantz MG; NICHD Pelvic Floor Disorders Network. Measuring Clinical Importance in a Trial of Interventions for Mixed Urinary Incontinence-Reply. JAMA. 2020 Feb;323(5):479–80.

22 Man-Son-Hing M, Laupacis A, O’Rourke K, Molnar FJ, Mahon J, Chan KB, et al. Determination of the clinical importance of study results. J Gen Intern Med. 2002 Jun;17(6):469–76.

23 Leichsenring F, Steinert C, Ioannidis JP. Toward a paradigm shift in treatment and research of mental disorders. Psychol Med. 2019 Oct;49(13):2111–7.

24 Fava GA, Carrozzino D, Lindberg L, Tomba E. The clinimetric approach to psychological assessment: a tribute to Per Bech, MD (1942–2018). Psychother Psychosom. 2018;87(6):321–6.