Psychotherapy Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tpsr20

Direct retrospective measurement of therapeutic changes: an example using the Czech version of the Questionnaire of Personal Changes (Q-PC)

Tomáš Řiháček, Kateřina Macková & Hynek Cígler

To cite this article: Tomáš Řiháček, Kateřina Macková & Hynek Cígler (08 Jul 2024): Direct retrospective measurement of therapeutic changes: an example using the Czech version of the Questionnaire of Personal Changes (Q-PC), Psychotherapy Research, DOI: 10.1080/10503307.2024.2370357

To link to this article: https://doi.org/10.1080/10503307.2024.2370357

Published online: 08 Jul 2024.

Psychotherapy Research, 2024
https://doi.org/10.1080/10503307.2024.2370357

Direct retrospective measurement of therapeutic changes: an example using the Czech version of the Questionnaire of Personal Changes (Q-PC)

TOMÁŠ ŘIHÁČEK, KATEŘINA MACKOVÁ, & HYNEK CÍGLER

Department of Psychology, Faculty of Social Studies, Masaryk University, Brno, Czech Republic

(Received 18 March 2024; revised 7 June 2024; accepted 16 June 2024)

Abstract

Objective: The study aimed to test the psychometric properties of the Czech translation of the Questionnaire of Personal Changes (Q-PC), a measure designed for retrospective (direct) measurement of change in psychotherapy.
Methods: A sample of group psychotherapy clients ($N = 222$) and a nonclinical sample ($N = 167$) sample were used. Clients in the clinical sample were administered the Q-PC in addition to several pre–post outcome measures. Confirmatory factor analysis, correlational analysis, and structural equation modeling were used to test the Q-PC's factor structure, longitudinal measurement invariance, reliability, convergent validity, sensitivity to change, and other psychometric properties.
Results: The Q-PC demonstrated a unidimensional structure that was strictly invariant between two follow-up measurement waves. The measure also demonstrated excellent reliability and sensitivity to change and good convergent validity. Furthermore, it demonstrated a similar relationship to baseline severity as the pre–post outcome measures.
Conclusions: The retrospective measurement of change is a promising approach that has the potential to complement the traditional pre–post measurement of change.

Keywords: Questionnaire of Personal Changes; retrospective measurement of change; direct measurement of change; factor analysis; sensitivity to change; positive change bias

Clinical or methodological significance of this article: This study investigated the applicability of retrospective (direct) measurement of change in psychotherapy. A retrospective measure (Questionnaire of Personal Changes) captured therapeutic change in a similar manner and yielded scores comparable to those obtained via traditional, pre–post outcome measurement. The study showed that retrospective change measurement is a promising approach with a potential to complement the traditional pre–post measurement approach.

Valid and reliable assessment of psychotherapy outcomes has been a challenge since the dawn of psychotherapy research. Many self-report outcome measures have been developed, some of which have become widely used and accepted as the field standard (e.g., Outcome Questionnaire-45; Lambert et al., 1996; Clinical Outcomes in Routine Evaluation-Outcome Measure, Evans et al., 2002; and Outcome Rating Scale, Miller et al., 2003). The assessment of the therapeutic change using these measures requires comparing pre- and post-treatment scores to derive a difference score. Because the change score is only indirectly derived from two measurements of the clients' momentary status, this is sometimes referred to as *indirect* measurement of change (Krampen, 2010a). In some situations, this approach may be infeasible, since pretreatment scores are not always available and obtaining them retrospectively comes with its own methodological challenges (Howard et al., 1981). This is why some authors

Correspondence concerning this article should be addressed to Tomáš Řiháček, Department of Psychology, Faculty of Social Studies, Masaryk University, Joštova 10, 602 00 Brno, Czech Republic. Email: rihacek@fss.muni.cz

coined the idea of *direct* assessment of change that is based on clients' one-time retrospective estimation of the amount of change that took place over a specified period, typically over the course of their treatment (Krampen, 2010a; Sandell & Wilczek, 2016). Several measures have been developed that allow clinicians to measure directly how much clients feel their problems have changed or to which degree they have attained their treatment goals. Examples include the Goal Attainment Scale (GAS; Kiresuk & Sherman, 1968) and its revision (GAS-R; Kiresuk & Lund, 1979), the Questionnaire to Assess Changes in Experiencing and Behavior (QCEB; Zielke & Kopf-Mehnert, 1978), the Bochum Change Questionnaire 2000 (BCQ 2000; Willutzki et al., 2013), and the Questionnaire of Personal Changes (Q-PC; Krampen, 2010a, 2010b). Despite their potential, these methods remain largely unexplored. Therefore, this study aimed to assess whether the direct approach to measurement (represented by the Q-PC in this study) is a viable option to the traditional indirect (i.e., pre–post assessment) approach. It should be noted that the "direct measurement" of change here means clients' perception of or "feelings about" the change they made. However imprecise this may be compared to the indirect assessment, perceptions of actual experience may, in fact, have higher clinical relevance for both clients and their therapists.

In addition to the practical advantage (i.e., no need for a pretreatment measurement), the direct measurement of change strives to solve the problem known as "response shift" in the psychometric literature (Vanier et al., 2021). If a person experiences a change during treatment, then we may assume that this change is not purely quantitative or incremental, as Sandell and Wilczek (2016) phrase it. Rather than merely decreasing clients' symptoms, psychotherapy may change the very frame of reference within which clients assess their symptoms and life. This shift in perspective may then cause pre- and posttreatment scores on a traditional outcome measure to no longer be comparable. For instance, clients may underestimate the severity of their problems before treatment or become less worried about their symptoms during treatment (Stänicke & McLeod, 2021). This problem is also known as measurement noninvariance in the psychometric literature, which occurs when two measurements differ in the factorial structure, factor loadings, intercepts, or residual item variances and thus the factor scores do not represent identical constructs (Fokkema et al., 2013). Fokkema et al. demonstrated that the Beck Depression Inventory, a frequently used outcome measure, was not invariant between pre- and post-treatment measurements. In such situations, the indirect, pre–post measurement of change would not correspond to clients' actual experience of change (Roubal et al., 2018). In contrast, the direct approach to change measurement allows clients to assess the amount of change, as meaningfully perceived in retrospect – i.e., from the perspective from which clients would evaluate their treatment success anyway.

Another problem related to the indirect measurement of change is regression to the mean. Irrespective of the therapeutic effect, the pre- and posttreatment measurements can be considered parallel measurements. The more extreme value clients report before the treatment, the more likely they are to report a value closer to the mean after the treatment (Hsu, 1995). This problem challenges the meaningfulness of comparing repeated measurements in psychotherapy (Nesselroade & Ghisletta, 2003).

However, direct measurement has its own problems. It relies on retrospection and, therefore, is expected to be prone to a variety of memory biases (Neusar, 2014; Rodgers & Elliott, 2015), although Flückiger et al. (2007) demonstrated that it was no more susceptible to mood effects than post-treatment measurements on usual pre–post measures. Clients may find it difficult to remember how they felt before the treatment and may overestimate their pretreatment distress level (Safer & Keuler, 2002). The direct measurement may also be more prone to the social desirability effect (Adams et al., 1999). Although this type of bias is known to operate in the context of pre–post assessment (known as the "hello-goodbye" effect; Elliott, 2002), a one-time retrospective assessment makes it easier for clients to overestimate the change to please their therapist. We refer to this as the positive change bias hypothesis. Thus, a direct measurement may thus not be more valid than an indirect measurement. Nevertheless, it represents a valid perspective on its own and should be explored alongside the more traditional, pre–post measurement approach.

Studies that investigated methods of direct change assessment concluded that they show medium to strong correlations with traditional, indirect methods. The correlation coefficient did not exceed the value of $r = .40$ in Michalak et al. (2003), $r = .52$ in Krampen (2010a), .60 in Sandell and Wilczek (2016), and $r = .72$ in Flückiger et al. (2007). This suggests that while the direct and indirect methods are related, they represent different perspectives and are thus not interchangeable. Rather, they provide a more complex picture of therapeutic change when combined. However, the correlations can also be affected by a lower reliability of the measures, and the true correlation between the latent variables may be higher.

In the context of indirect change measurement, it is possible to compare the pre- and post-treatment measurements and study the relationship between them. For instance, it is assumed that change scores are usually negatively correlated with clients' pretreatment status on the given variable (Castonguay et al., 2021; Chiou & Spreng, 1996). This means that clients with more severe baseline distress tend to demonstrate greater change during treatment, while clients who are better off at baseline show smaller changes. In the context of direct change measurement, it is impossible to address this question if we only have a single score (i.e., a one-time direct estimate of perceived change). Nevertheless, if the direct measurement approach is to be considered commensurable to the pre–post approach, it should demonstrate a similar pattern of relationships. In previous studies, direct measurement was only negligibly to weakly related to baseline distress (Flückiger et al., 2007; Michalak et al., 2003; Sandell & Wilczek, 2016). Therefore, in our study, we employed both a direct and an indirect measurement of change to address this problem empirically.

To allow for direct quantitative change measurement, Krampen (2010a, 2010b) developed the Questionnaire of Personal Changes (Q-PC). The Q-PC asks clients to assess the extent of change, either positive or negative, experienced during a treatment. The questionnaire was constructed as a brief, 12-item unidimensional measure. The items represent perceived changes in behavior (six items) and experience (six items). Furthermore, pairs of items represent six more specific areas of functioning (i.e., relaxation; emotional stabilization; self-regulation; utilization of one's own abilities and performance; well-being and coping with difficulties; and self-efficacy and control). The items are not disorder-specific and can all be framed as aspects of self-efficacy and coping. Overall, they indicate an improved, unchanged, or deteriorated ability to cope with life problems and situations that were previously perceived as difficult or problematic. Theoretically, the measure is based on Grawe's integrative approach to psychological therapy (Grawe, 2004).

To date, the psychometric properties of the Q-PC have only been reported in the original study, which has been published in both English (Krampen, 2010a) and German (Krampen, 2010b). Although an exploratory factor analysis conducted in the original study suggested the existence of two or three factors, Krampen (2010a) did not interpret them and treated the scale as unidimensional. Cronbach's alpha of the total score varied between .91 and .96 in the original study, supporting this decision. However, Cronbach's alpha is not a measure of dimensionality (Schmitt, 1996); therefore, the dimensionality of the measure remains to be tested. Krampen also documented convergent and divergent validity of the scale. The author also demonstrated the scale's sensitivity to change by comparing psychotherapy clients' scores to those reported by a waiting-list control group. The measure has been also used in another study that reported a Cronbach's alpha of .94 but did not investigate the factor structure (Masroor et al., 2013).

Aim of Study

This study aimed to test the psychometric properties of the Czech version of the Questionnaire of Personal Changes (Q-PC). We used data from a study on the effectiveness of multicomponent group-based treatment (Pourová et al., 2024; Řiháček et al., 2022), in which the Q-PC was administered alongside several pre–post outcome measures targeting depression, anxiety, and well-being. While findings based on the pre–post measures were published in the abovementioned studies, the Q-PC data are presented in the current study for the first time.

First, we assessed the basic psychometric properties of the Czech Q-PC, including: (a) the factor structure (we expected a unidimensional structure), (b) longitudinal measurement invariance (between two waves of follow-up change assessment), and (c) reliability of the Q-PC score. Although the measure was conceived as unidimensional (Model 1), we also tested two alternative models inherently present in the measure construction: a two-factor model with the behavioral and experiential factors (Model 2), and a unidimensional model with freed residual correlations for each of the six item pairs. Second, we tested for (d) the possibility that the Q-PC overestimates change (the positive change bias hypothesis). For this purpose, we used a nonclinical sample collected specifically for this study. Third, we used the clinical sample's pre–post data on depression, anxiety, and well-being to (e) assess the Q-PC's convergent validity with pre–post assessment of change on the three outcome measures, (f) assess the Q-PC's sensitivity to change (compared to change scores of the three pre–post outcome measures), and (g) test the degree to which the Q-PC scores reflect clients' pre-treatment status (compared to the three pre–post outcome measures).

Method

Study Design and Sample

The study was approved by the Research Ethics Committee of the Masaryk University (ref. no. EKV-2017-029-R1).

Table 1. Sample characteristics.

	Clinical sample (6-month follow-up)	Clinical sample (12-month follow-up)	Nonclinical sample
N	222	190	167
Demographic information			
Age (years)			
Mean (SD)	39.5 (11.6)[a]		

Clinical sample. We used data from an uncontrolled naturalistic study on group psychotherapy effectiveness collected across seven clinical sites in the Czech Republic (Pourová et al., 2024; Řiháček et al., 2022). The treatment length varied between four and twelve weeks, with the most common length being six weeks. At five sites, clients received five sessions of face-to-face group psychotherapy per week, while at two sites, they received only three or four sessions. Typically, a session lasted 90 min (except for one site, where sessions lasted 75 min). The treatment was non-manualized, mostly psychodynamic, with the integration of humanistic and experiential approaches. Group psychotherapy sessions were supplemented with other activities such as art therapy, relaxation training, music therapy, and others, depending on the site.

Of the total of 736 clients, 444 agreed to participate in the study. The clients were administered the PHQ-9, GAD-7, and WHO-5 at baseline (paper-and-pencil), at treatment termination (paper-and-pencil), and at six- and 12-month follow-ups (online). The Q-PC was administered only at the six- and 12-month follow-up surveys, which were answered by 222 and 190 clients, respectively. Only these clients were included in the study. See Table 1 for the sample description.

Nonclinical *sample*. The nonclinical sample was recruited via social media networks. Only participants who met the following criteria were included: (a) age 18 or older; (b) no psychiatric diagnosis in the last 12 months; and (c) no use of any psychological, psychotherapeutic, or psychiatric services in the last 12 months. Of the 235 participants who opened the survey, 202 were eligible, and of those, only 167 completed the survey. The participants answered an anonymous one-time online survey that contained the Q-PC and WHO-5 and, therefore, responded to the Q-PC only once. See Table 1 for the sample description.

Measures

Questionnaire of personal changes (Q-PC). The Q-PC (Krampen, 2010a, 2010b) is a 12-item self-report measure designed for "direct" retrospective measurement of therapeutic change. Items 1, 2, 6, 7, 8, and 9 measure aspects of behavior, while Items 3, 4, 5, 10, 11, and 12 are focused on experience. More specifically, Items 1 and 2 focus on psychological and physical relaxation, Items 4 and 5 on emotional stabilization, Items 6 and 7 on self-regulation, Items 8 and 9 on utilization of one's own abilities and performance, Items 3 and 11 on well-being and coping with difficulties, and Items 10 and 12 on self-efficacy and control (Krampen, 2010b). Clients rated each item on a seven-point bipolar scale ranging from +3 = strong positive change to –3 = strong negative change with the midpoint labelled as "no change." They are asked to "think back to the time prior to the beginning their treatment" and assess the extent of change over the whole course of the treatment. Thus, clients in the clinical sample were asked to rate the overall perceived change from the beginning of the treatment at both follow-up measurements. Clients in the nonclinical sample had no therapy and were asked to think back over the last six months of their life.

The scale was translated into Czech from the English version. Five independent Czech translations were made by native Czech speakers (a psychology student, two psychologists, and two laypeople). Second, all translations were discussed by a group of three people (the two psychologists and the psychology student) and consolidated into a single version. Third, this version was back-translated into English by a bilingual, native English speaker and compared to the original English version. Fourth, the back-translation was discussed with the author of the scale, and minor corrections were made based on this discussion. Fifth, the final Czech version was field-tested with five respondents to check the comprehensibility of the items.

Patient health questionnaire-9 (PHQ-9). The PHQ-9 (Kroenke et al., 2001) is a nine-item self-report measure for screening the severity of depressive symptoms over the past two weeks. Clients rate each item on a four-point Likert-type scale where 0 means "not at all" and 3 means "nearly every day." The scale has been validated in the Czech Republic (Daňsová et al., 2016). In this study, the Cronbach's α at baseline was α = .81.

Generalized anxiety disorder screener (GAD-7). The GAD-7 (Löwe et al., 2008) is a seven-item self-report measure of anxiety symptoms over the last two weeks. Clients rate each item on a four-point Likert-type scale where 0 means "not at all" and 3 means "nearly every day." The scale has been validated in the Czech Republic (Prikner, 2021). In this study, the Cronbach's α at baseline was α = .86.

Well-being index (WHO-5). The WHO-5 (Bech et al., 2003) is a self-report measure of well-being operationalized as positive affect (Kusier & Folker, 2020). The scale consists of five items (four assessing hedonia, one item assessing eudaimonia), and each item is rated on a six-point Likert-type scale where 5 means "all of the time" and 0 means "at no time." In a systematic review, Topp et al. (2015) demonstrated that the measure has good psychometric properties, including clinimetric validity, sensitivity, and specificity, across many studies. The Czech version has not yet been validated. In this study, the Cronbach's α at baseline was α = .85.

Demographic questionnaire. The demographic questionnaire contained questions about the respondents' gender, age, and education. Furthermore, participants in the nonclinical sample were asked about their mental health status to determine their eligibility.

Statistical Analysis

The statistical analysis was conducted using R software version 4.2.3 (R Core Team, 2023). First, we conducted a confirmatory factor analysis (CFA) to test the suggested models using the lavaan (Rosseel, 2012) and semTools (Jorgensen et al., 2022) packages. Since the values of some Q-PC items were nonnormally distributed, we used the robust maximum likelihood estimator (MLR). The model was defined as congeneric, and the variance of the latent variable was set to 1. The model fit was assessed using the standardized root mean square residual (SRMR), the root mean square error of approximation (RMSEA), and the Tucker-Lewis index (TLI). Robust chi-squared statistics and their degrees of freedom were reported per convention. However, we did not interpret them because this test is sensitive to the sample size (leading to a higher likelihood of rejecting a model in large samples) and nonnormal data distribution. Hu and Bentler (1999) recommended values close to 0.08 for the SRMR, 0.06 for the RMSEA, and 0.95 for the TLI as cutoffs for a fitting solution. Other authors, however, have suggested less stringent criteria for model rejection, i.e., RMSEA > 0.10 and TLI < 0.90 (Brown, 2015).

Second, we tested the measurement invariance between the clinical (six-month follow-up) and nonclinical samples and longitudinal measurement invariance between the two waves of data collection (i.e., six- and 12-month follow-up) within the clinical sample. We gradually constrained the factor loadings (metric invariance), item intercepts (scalar invariance), and residual variances (strict invariance). In the case of longitudinal invariance, the measurement errors were allowed to freely covary across measurement waves. We did not constrain means since equality of means was expected in neither comparison. The invariance was assessed by a change in fit compared to a previous model; a change in TLI ≥ 0.005 (for all levels of invariance), supplemented by a change in RMSEA ≥ 0.010 (for all levels of invariance) or a change in SRMR ≥ 0.025 (for metric invariance) and ≥ 0.005 (for scalar and strict invariance) indicate noninvariance in samples with $N <$ 300 (Chen, 2007).

Third, to assess the reliability of the Q-PC score, we reported the alpha and omega coefficients. To facilitate change assessment, we also calculated the reliable change index (RCI, Jacobson & Truax, 1991). The standard RCI formula (see Formula 1) accounts for the fact that a difference score contains two measurement errors (i.e., the pre- and post-measurement). However, in the case of Q-PC, only one measurement is employed, and therefore, the $\sqrt{2}$ term is omitted from the formula (Formula 2), making the formula identical to that of the standard error of measurement (Harvill, 1991). Hence, we used Formula 2 to calculate RCI for the Q-PC and Formula 1 to calculate RCI for the remaining instruments (i.e., PHQ-9, GAD-7, and WHO-5).

$$RCI = 1.96 * \sqrt{2} * SD * \sqrt{(1 - REL)} \tag{1}$$

$$RCI = 1.96 * SD * \sqrt{(1 - REL)} \tag{2}$$

where SD is the standard deviation and REL is the reliability of the measure (we used Cronbach's alpha in our study).

Fourth, to test the positive change bias hypothesis, we estimated the amount of change reported by people who, on average, were not expected to report any change (i.e., the nonclinical sample). We computed the mean change and tested it against the null hypothesis of no change using a one-sided t-test.

Fifth, we assessed the Q-PC's convergent validity with pre–post assessment of change on the PHQ-9, GAD-7, and WHO-5. For this purpose, we calculated mean total scores for Q-PC (at six- and 12-month follow-up) and for PHQ-9, GAD-7, and WHO-5 (at baseline and both follow-up measurements). We only did so for clients who responded to at least 80% of items on a measure; otherwise, we treated the total score as missing. Afterwards, we subtracted the PHQ-9, GAD-7, and WHO-5 follow-up scores from the baseline to obtain change scores. Finally, we correlated the change scores to each other and to the Q-PC score. Since some of the total scores tended to be non-normally distributed, we used Spearman's correlation coefficient (r_s). For documentation purposes, we also reported the reliability of the difference scores. In the case of the Q-PC, the reliability of the difference score was equal to Cronbach's alpha because the Q-PC score itself was considered a difference score. In the case of the pre–post measures, the reliability of the difference score was determined using the following formula (Williams & Zimmerman, 1977):

$$\rho = \frac{\frac{rel(x) - rel(y)}{2} - cor(x, y)}{1 - cor(x, y)} \tag{3}$$

where *rel(x)* and *rel(y)* are the reliabilities of measurement X (baseline) and Y (follow-up) and *cor(x,y)* is the correlation between the two measurements. We used Cronbach's alpha for *rel* and Pearson correlation for *cor*.

Another approach to assess convergent validity was to explore whether the Q-PC classified clients into improved versus unchanged/deteriorated in the same manner as the PHQ-9, GAD-7, and WHO-5 difference scores. For this purpose, we recoded each measure's change scores into 1 (for clients who improved) and 0 (for clients who did not change or deteriorated). As recommended by Kottner et al. (2011), we then assessed pairwise agreement using both Cohen's kappa and the percentage of identically classified cases.

Sixth, we compared the Q-PC's sensitivity to change to that of the pre–post outcome measures. To do so, we computed Cohen's *d* by dividing the change score (i.e., the Q-PC raw score and the difference scores of the pre–post outcome measures) by the standard deviation of the change score and compared these standardized effect sizes across measures. Furthermore, we used the RCI concept to classify clients into those reliably improved (i.e., improvement larger than the RCI), those reliably deteriorated (i.e., deterioration larger than the RCI), and those without any reliable change (i.e., zero change or improvement/deterioration smaller than the RCI). We then compared the measures' ability to detect reliable change.

Seventh, to estimate the extent to which the Q-PC reflects clients' baseline severity, we employed structural equation modeling (SEM). The primary aim of this analysis was to decompose the Q-PC variance explained by clients' baseline status (represented by the baseline measurement in the context of indirect measurement) from the variance explained by change (represented by the difference score in the context of indirect measurement). We defined three latent variables, namely, baseline status, six-month follow-up change, and 12-month follow-up change. The *baseline status* latent variable loaded onto all baseline measures (i.e., PHQ-9, GAD-7, and WHO-5), all six-month follow-up measures (i.e., PHQ-9, GAD-7, WHO-5, and Q-PC), and all 12-month follow-up measures (i.e., PHQ-9, GAD-7, WHO-5, and Q-PC). Furthermore, the *six-month follow-up change* latent variable loaded onto all six-month follow-up measurements, and analogically, the *12-month follow-up change* latent variable loaded on all six-month follow-up measurements. By holding both *follow-up change* latent variables orthogonal to the baseline status, we ensured that the *baseline status* variable "drained" the portions of the follow-up measurement variances explained by the baseline severity. Consequently, both *follow-up change* latent variables can be interpreted as clients' latent difference scores. The *follow-up change* latent variables were set to freely covary, as were the residuals of each observed variable across time. To ensure longitudinal measurement invariance, we added several constraints to the model. First, to ensure metric invariance, we fixed each observed variable's loadings to the same value within the *baseline status* latent variable (e.g., the baseline, six-month, and 12-month PHQ-9 loadings were set to have the same value within the *baseline status*). Furthermore, we allowed the loadings to change in both *follow-up change* latent variables but fixed the proportion of change to the same value for all observed variables within the latent variable (the so-called proportional constrains). Second, we constrained each observed variable's intercept to be the same across measurements (scalar invariance). We used the robust maximum likelihood estimator (MLR) to estimate the model and the full information maximum likelihood (FIML) method to treat missing data. Both the latent and observed variables were standardized. Due to the small sample size, we did not model each measure's total score itself as a latent variable. Instead, we calculated sum scores and treated them as observed variables in the model.

Results

Descriptive Statistics

The descriptive statistics of the Q-PC items are reported in Table 2. The distribution of most Q-PC items was positively skewed in the clinical sample, but none indicated a ceiling effect. In the six-month follow-up clinical sample, 27 clients did not answer the Q-PC. In the 12-month follow-up clinical sample, 11 clients did not answer the Q-PC. In the nonclinical sample, six clients did not answer the Q-PC. This left us with $N = 195$ in the six-month follow-up, $N = 179$ in the 12-month follow-up, and $N = 161$ in the nonclinical sample.

Confirmatory Factor Analysis

We started with testing factor models in the six-month follow-up clinical sample. First, we tested the unidimensional model (Model 1), which had unacceptable fit (see Table 3 for fit indices). Second, we tested a two-factor model composed of the behavioral and experiential factors (Model 2). However, fit increased only negligibly because the two factors were almost perfectly correlated ($r = .97$). Third, we tested the unidimensional model and freed residual covariances between pairs representing the same area of functioning. The fit increased considerably, and although RMSEA was still not optimal (0.095), it was acceptable using the less stringent criteria (Brown, 2015). Kenny et al. (2015) also argued that, with small samples, the RMSEA often falsely indicates a poor fitting and should not be used as a sole criterion to reject a model. Because exploratory analyses did not reveal any other meaningful solution, we accepted Model 3 as our final model. See Figure 1 for the model parameters at the six-month follow-up and Supplement 1 for the tabulated parameters of all three samples.

Measurement Invariance

Model 3 was strictly invariant between the two measurement waves in the clinical sample. However, only configural invariance was achieved between the clinical (six-month follow-up) and nonclinical samples. There was a notable drop in fit between the configural and metric invariant models, suggesting that the differences in item loadings were too sizeable between the two samples (see Supplement 1). See Table 4 for fit indices.

Reliability

The measurement reliability was $\alpha = .95$ ($\omega = .93$) for the six-month follow-up in the clinical sample, $\alpha = .95$ ($\omega = .94$) for the 12-month follow-up in the clinical sample, and $\alpha = .91$ ($\omega = .89$) for the nonclinical sample. The reliable change index based on the alpha coefficient was RCI = 5.62 and 5.51 for the six- and 12-month follow-up measurements, respectively.

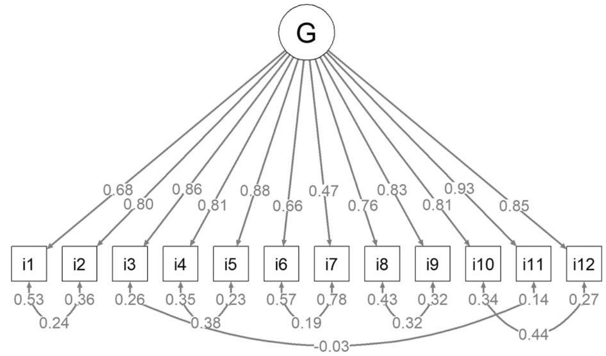

Figure 1. The final model (Model 3) at the six-month follow-up (parameters completely standardized).

Positive Change Bias Hypothesis

Respondents in the nonclinical sample (i.e., those who were not expected, on average, to change) reported a mean change of $M = 2.95$ ($SD = 11.29$), which was statistically significantly different from zero, $t(158) = 3.29$, $p < .001$. This translated to Cohen's $d = 0.26$ with 95% CI [0.11, 0.41].

Convergent Validity

Table 5 shows that there were substantial correlations between the Q-PC scores and the difference scores on the pre–post outcome measures (r_s between .42 and .60), although the pre–post outcome measures correlated with each other more strongly (r_s between .62 and .75). A similar conclusion can be drawn from the percentage of agreement on the improved versus unchanged/ deteriorated status: the agreement between Q-PC and pre–post change scores ranged from 71% to 80%, while the agreement between the pre–post change scores themselves ranged from 77% to 85%. The findings were consistent between the six- and 12-month follow-up measurements. Notably, the kappa coefficients suggested considerably lower levels of agreement compared to the percentages because Cohen's kappa penalizes for the imbalance in the proportion of the variable levels (Feinstein & Cicchetti, 1990). In our sample, improvement was substantially more frequent than no change/deterioration (see Table 6).

Sensitivity to Change

The standardized effect sizes (Cohen's d) showing clients' change between the baseline and the six- or 12-month follow-up measurement were larger for the Q-PC (see Table 6). However, as we reported above, the Q-PC appears to introduce positive change bias. Therefore, we repeated the same

Table 2. Q-PC item descriptives.

	Items	Clinical 6-month follow-up		Clinical 12-month follow-up		Non-clinical	
		M	SD	M	SD	M	SD
1	I can relax much better. *Dokáži mnohem lépe odpočívat.*	4.18	1.13	4.34	1.07	0.41	1.38
2	I can unwind better and take it easy. *Dokáži se lépe uvolnit a brát věci s nadhledem.*	4.27	1.12	4.39	1.08	0.62	1.44
3	Overall I feel healthier. *Celkově se cítím zdravější.*	3.91	1.43	4.13	1.35	0.32	1.45
4	I feel less anxious thinking about the future. *Méně se obávám budoucnosti.*	3.86	1.37	3.99	1.44	−0.18	1.47
5	I feel calmer and more well-balanced. *Cítím se klidnější a vyrovnanější.*	3.96	1.33	4.18	1.37	0.26	1.43
6	I sleep better. *Lépe spím.*	3.59	1.29	3.70	1.30	−0.05	1.46
7	I take less medication. *Užívám méně léků.*	3.18	1.49	3.39	1.56	0.08	0.62
8	I have more stamina and do not give up as easily. *Mám větší výdrž a nevzdávám se tak snadno.*	3.77	1.36	3.97	1.39	0.43	1.29
9	I can concentrate much better. *Dokáži se mnohem lépe soustředit.*	3.60	1.31	3.86	1.33	0.05	1.29
10	I cope with unexpected events more easily. *Snadněji zvládám nečekané události.*	3.88	1.30	3.96	1.29	0.43	1.28
11	I feel better. *Cítím se lépe.*	4.15	1.42	4.28	1.41	0.44	1.32
12	I deal with stress and pressure better. *Lépe se vyrovnávám se stresem a situacemi, kdy jsem pod tlakem.*	3.83	1.32	4.01	1.33	0.14	1.43

Note: Czech translation in italics.

analysis for Q-PC scores diminished by the mean change reported by the nonclinical sample. After this correction, the Q-PC effect size dropped considerably but remained comparable to or higher than that of the WHO-5 (see Table 6).

In terms of the measures' ability to detect statistically reliable changes, the Q-PC outperformed the remaining measures by a large margin, both in terms of improvement and deterioration. This lead was preserved even after controlling for the positive change bias. Again, the results were largely consistent across the six- and 12-month follow-up measurements.

Relationship of the Q-PC Score to Baseline Severity

We estimated an SEM to disentangle the proportion of Q-PC score variance explained by clients' baseline status from that explained by perceived change. The model is displayed in Figure 2 (note that only the parameters of interest are displayed in Figure 2; see Supplement 3 for the full list of parameters). The model fit was excellent, $\chi^2(37) = 51.63$, $p = .056$, SRMR = 0.038, RMSEA = 0.046 [0.000, 0.078], TLI = 0.988. As expected, the *baseline status* factor loadings were highest for the baseline measurement variables

Table 3. Fit indices for confirmatory factor analysis in the clinical sample.

Invariance	χ^2	*df*	BIC	SRMR	RMSEA	TLI
Clinical sample (six-month follow-up, N = 195)						
Model 1 (unidimensional)	177.6***	54	6040	0.049	0.127	0.897
Model 2 (two-factor)	173.2***	53	6038	0.047	0.127	0.898
Model 3 (unidimensional with residual cov.)	111.6***	48	5977	0.040	0.095	0.943
Clinical sample (12-month follow-up, N = 179)						
Model 3	112.4***	48	5595	0.037	0.094	0.943
Nonclinical sample (N = 161)						
Model 3	68.2*	48	5526	0.044	0.058	0.960

Note: SRMR = standardized root mean square residual, RMSEA = robust root mean square error of approximation, TLI = robust Tucker-Lewis index. * $p < .05$, *** $p < .001$.

Table 4. Fit indices for invariance testing.

Invariance	χ^2	*df*	$\Delta\chi^2$	*Δdf*	BIC	ΔBIC	SRMR	*Δ*SRMR	RMSEA	*Δ*RMSEA	TLI	*Δ*TLI
Clinical (six-month follow-up) vs. *nonclinical*												
Configural	180.5***	96			11686		0.039		0.081		0.948	
Metric	218.1***	107	41.1***	11	11664	−22	0.087	0.048	0.087	0.006	0.940	−0.008
Scalar	248.0***	118	31.3***	11	11632	−32	0.092	0.005	0.089	0.002	0.937	−0.003
Strict	425.9***	130	152.4***	12	11792	160	0.203	0.111	0.129	0.040	0.868	−0.069
Six-month vs. *12-month follow-up in the clinical sample*												
Configural	349.7***	227			8705		0.042		0.065		0.946	
Metric	357.7***	238	8.0	11	8660	−45	0.051	0.009	0.063	−0.002	0.950	0.004
Scalar	368.7***	249	10.4	11	8616	−44	0.051	0.001	0.061	−0.002	0.952	0.002
Strict	377.9***	261	10.4	12	8570	−46	0.051	0.000	0.060	−0.002	0.955	0.003

Note: BIC = Bayesian information criterion, SRMR = standardized root mean square residual, RMSEA = robust root mean square error of approximation, TLI = robust Tucker-Lewis index. *** $p < .001$.

and gradually decreased with the six- and 12-month follow-up measurements.

The parameters of interest are the Q-PC loadings and similarity (or lack thereof) to the pre–post measures' loadings. To facilitate the interpretation, we will now refer to their absolute values (all signs were in the expected direction, and they differed because in some of them, a higher score represented higher distress, while in others, a higher score represented lower distress). In the case of the six-month follow-up measurement, the *baseline status* loading on the Q-PC was $|\lambda| = .36$ (compared to the pre–post measures' loadings of $|\lambda|$ between .39 and .45). This means that the Q-PC score was influenced by the baseline status to a slightly lower degree compared to the pre–post measures. The *six-month follow-up change* factor had a $|\lambda| = .62$ loading on the Q-PC (compared to the pre–post measures' loadings of $|\lambda|$ between .73 and .84). Again, the magnitude of the loading was slightly lower for the Q-PC. A similar pattern was demonstrated in the case of the *12-month follow-up change* factor, essentially replicating the six-month follow-up findings.

Discussion

The aim of this study was twofold: (1) to test the psychometric properties of the Czech translation of the Q-PC and, using this exemplar, (2) to draw more general conclusions about the applicability of retrospective (direct) measurement of change in psychotherapy. We used data from an existing study where the Q-PC was administered alongside a set of traditional pre–post outcome measures. The Q-PC was administered twice (at the six-month and 12-month follow-ups), and we used these two measurement waves as a means of internal cross-validation. The two waves yielded comparable results across all analyses, supporting the validity of this study's findings.

In terms of the factor structure, the Q-PC can be considered a unidimensional measure. Although the initial unidimensional model did not fit the data, the fit increased considerably after residual covariances between pairs of items representing the same area of functioning were allowed to covary. The alternative two-factor structure with the behavioral and experiential factors was not tenable, and the extremely high correlation between the two factors suggested that the two factors cannot be empirically distinguished from each other. The modified unidimensional model was strictly invariant between the two measurement waves in the clinical sample, which means that the Q-PC can be safely used to compare the magnitude of change between different posttreatment and follow-up measurement points. However, since we only compared six- and 12-month follow-up measurements, we cannot generalize this finding to substantially longer periods (i.e., multiples of years). The reliability of the Q-PC total score was excellent in the clinical sample, echoing Krampen's (2010a) original findings.

We hypothesized that in the context of retrospective measurement, clients may be inclined to report positive change even if there was none (we called this positive change bias). Our data from the nonclinical sample confirmed this hypothesis: people who, on average, were not expected to change reported a small to medium change using the Q-PC. There are several possible explanations, including a tendency to overestimate the initial distress level (Safer & Keuler, 2002), the social desirability effect (Adams et al., 1999), or the "hello-goodbye" phenomenon (Elliott, 2002). However, since our study is likely the first to explore the positive change bias in the context of retrospective change assessment, replication studies are needed to confirm this finding and to establish a more precise estimate of the reported change. Once this bias is examined more thoroughly, the mean change in the nonclinical sample can be used to correct for this

Table 5. Convergent validity.

	Six-month follow-up				12-month follow-up			
	Q-PC	ΔPHQ-9	ΔGAD-7	ΔWHO-5	Q-PC	ΔPHQ-9	ΔGAD-7	ΔWHO-5
Q-PC	ρ = .95	r_s = .45	r_s = .39	r_s = -.58	ρ = .95	r_s = .42	r_s = .45	r_s = -.60
ΔPHQ-9	κ = .29 76%	ρ = .76	r_s = .73	r_s = -.68	κ = .20 77%	ρ = .78	r_s = .75	r_s = -.69
ΔGAD-7	κ = .29 75%	κ = .60 85%	ρ = .81	r_s = -.62	κ = .41 82%	κ = .51 84%	ρ = .83	r_s = -.66
ΔWHO-5	κ = .26 71%	κ = .51 79%	κ = .50 79%	ρ = .84	κ = .44 80%	κ = .53 83%	κ = .49 80%	ρ = .86

Note: Δ = difference score (i.e., a difference from the baseline). Values on the diagonal represent the reliability of the difference score. Above diagonal values are Spearman correlations (all relationships were in the expected direction; negative signs in the WHO-5 columns reflect the fact that in WHO-5, a negative change score represented improvement). See Supplement 2 for raw variable correlations. Below diagonal values are Cohen's kappas and percentage of agreement.

bias. The most straightforward method is to use this mean change as a constant that is subtracted from the change reported by clients in clinical samples, as we did in our change sensitivity analysis. However, it should be noted that the Q-PC demonstrated only configural invariance between the clinical and nonclinical samples; therefore, any comparisons between clinical and nonclinical samples require caution. Nevertheless, a retrospective measurement of change in the context of a nonclinical sample in which, on average, no true change is expected is not a scenario in which the Q-PC would be routinely used. Therefore, the lack of strict invariance does not hinder the use of the Q-PC in clinical practice.

To assess convergent validity with the traditional, indirect change assessment, we correlated the Q-PC scores to difference scores of several pre–post outcome measures. The Q-PC yielded medium to large correlations with the traditional outcome measures. The highest correlations were with the WHO-5 well-being measure (34% to 36% shared variance), compared to correlations with the PHQ-9 depression and the GAD-7 anxiety measures (15% to 20% shared variance). This suggests that clients tended to treat the Q-PC as a measure of well-being rather than a distress measure. It may also mean that when clients are asked to assess "change," they implicitly interpret it as "change for the better." However small the proportions of shared variance may seem, they must be interpreted in the light of the shared variance between the pre–post measures themselves, which ranged between 38% and 56%. From this perspective, the Q-PC performed considerably well. Furthermore, the Q-PC also demonstrated a substantial ability to distinguish between improved and unchanged/deteriorated cases in a manner similar to the pre–post measures.

The Q-PC demonstrated excellent sensitivity to change. By having a smaller RCI, it was able to classify more clients as reliably improved or deteriorated compared to the pre–post outcome measures. The Q-PC's RCI was smaller than that of the pre–post measures for two reasons. First, the Q-PC had higher reliability than all pre–post measures. Second, the RCI formula for the pre–post outcome measures accounted for two measurement errors (i.e., the pre- and post-measurement), while the Q-PC's RCI formula only accounted for a single measurement error. In terms of the standardized effect size (Cohen's *d*), the Q-PC yielded the highest effects among all measures. Although this may reflect the positive change bias, the effect remained comparable to the pre–post measures (especially to the WHO-5) even after controlling for this bias.

Table 6. Sensitivity to change.

	Six-month follow-up				12-month follow-up			
	d	+	0	-	*d*	+	0	-
Q-PC	0.80	70%	19%	11%	0.94	73%	19%	8%
Q-PC (corrected)	0.57	59%	28%	13%	0.71	66%	23%	11%
ΔPHQ-9	0.73	31%	67%	2%	0.90	36%	62%	2%
ΔGAD-7	0.73	38%	57%	5%	0.79	31%	66%	3%
ΔWHO-5	0.56	30%	65%	5%	0.59	32%	64%	4%

Note: d = Cohen's *d*, + represents reliably improved cases, 0 represents cases without any reliable change, and – represents reliably deteriorated cases. Q-PC (corrected) shows the results for the Q-PC score corrected for the mean change of the nonclinical sample.

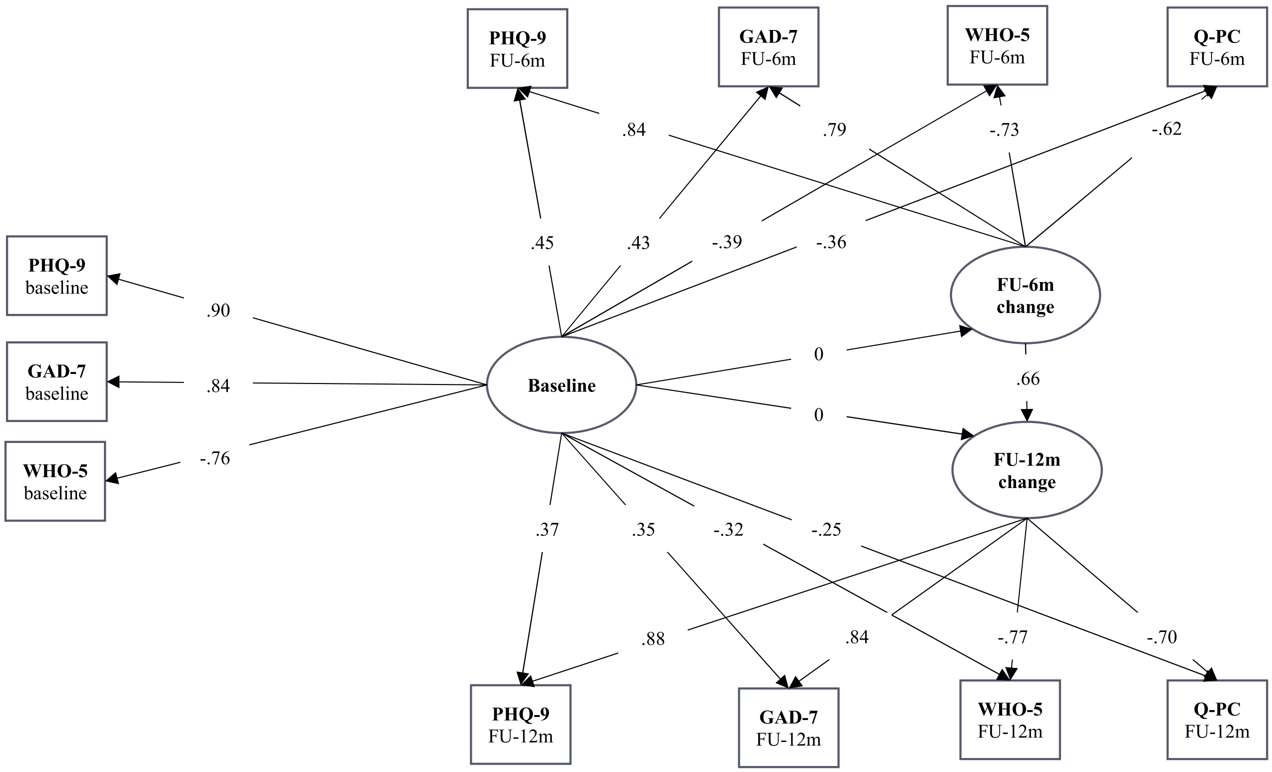

Figure 2. The structural equation model (completely standardized, residual variances and covariances omitted).

Finally, the Q-PC scores demonstrated a relationship to clients' baseline severity that was similar to how the pre–post difference scores are related to the baseline. This suggests that even in the context of a direct, retrospective measurement, clients can reliably compare their current status to their pre-treatment status and assess the magnitude of change in a way that is similar to the pre–post change measurement. It also means that the same psychometric effect that can be found in the context of pre–post measurement (i.e., the higher the baseline severity, the more room for improvement; Chiou & Spreng, 1996) applies to retrospective measurement. This feature is important because it makes retrospective and pre–post change scores directly comparable.

Limitations

The study was based on relatively small samples, which negatively influenced the reliability of the estimates. This problem was even more pronounced in the 12-month follow-up sample because of attrition. Moreover, we used total scores as manifest variables in the model. Larger samples would allow us to model total scores as latent variables, which may have yielded different results.

Furthermore, in the SEM, we treated baseline severity as a single latent factor, loading on the individual outcome variables (i.e., depression, anxiety, and well-being). This approach was supported by the recent concept of the general psychopathology factor (Caspi et al., 2014) and, empirically, by the high loadings of the baseline severity factor in our model. Nevertheless, this approach did not allow us to examine potential differences among various facets of psychopathology.

The generalizability of our findings is limited by the specificity of our sample. Although the sample was relatively heterogeneous in terms of clients' age, gender, diagnosis, and clinical sites, it only represented relatively short-term (four to 12 weeks) daily based group psychotherapy. Future studies are needed to examine the Q-PC features in other contexts, including different settings (e.g., individual psychotherapy), treatment length (e.g., long-term treatments), and time that elapsed between the treatment termination and the measurement (e.g., a few weeks vs. several years).

For the sake of simplicity, we assumed that clients in the clinical sample would report psychotherapy-related change, while participants in the nonclinical sample would, on average, report no change. However, change reported in both groups could have been influenced by extra-therapeutic life events. For instance, positive life events have been found to be associated with higher well-being (McCullough et al., 2000), positive affect (Clark & Watson, 1988), and happiness (Lyubomirsky et al., 2005). To control for the effect of these circumstances, life event would have to be recorded and factored into the analysis.

Conclusions

This study aimed to test the psychometric properties of the Czech translation of the Q-PC and, based on this, to draw more general conclusions about the applicability of retrospective (direct) measurement of change in psychotherapy. The Q-PC proved to be a valid and reliable retrospective measure of change in psychotherapy. A comparison of the Q-PC to several traditional, pre–post outcome measures suggested that it captures change in a similar manner and yields scores comparable to those obtained via indirect, pre–post outcome measurement. Additional studies are needed to replicate our findings under different conditions, including diverse therapeutic settings, types of treatment, treatment length, and distances between treatment termination and measurement. Our study showed that retrospective change measurement is a promising approach that, if supported by future studies, has the potential to complement and in some cases perhaps even replace the traditional pre–post measurement approach, especially if positive change bias is accounted for.

Funding

The study is from the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Data Availability

The data and the R code are available at https://osf.io/dfrma/

Supplemental Data

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10503307.2024.2370357.

ORCID

TOMÁŠ. ŘIHÁČEK http://orcid.org/0000-0001-5893-9289
HYNEK CÍGLER http://orcid.org/0000-0001-9959-6227

References

Adams, A. S., Soumerai, S. B., Lomas, J., & Ross-Degnan, D. (1999). Evidence of self-report bias in assessing adherence to guidelines. *International Journal for Quality in Health Care*, *11*(3), 187–192. https://doi.org/10.1093/intqhc/11.3.187

Bech, P., Olsen, L., Kjoller, M., & Rasmussen, N. (2003). Measuring well-being rather than the absence of distress symptoms: A comparison of the SF-36 mental health subscale and the WHO-Five well-being scale. *International Journal of Methods in Psychiatric Research*, *12*(2), 85–91. https://doi.org/10.1002/mpr.145

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Castonguay, L. G., Barkham, M., Youn, S. J., & Page, A. C. (2021). Practice-based evidence – findings from routine clinical settings. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garffield's handbook of psychotherapy and behavior change* (7th ed., pp. 191–222). Wiley.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chiou, J.-S., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *The Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, *9*, 158–167.

Clark, L. A., & Watson, D. (1988). Mood and the mundane: Relations between daily life events and self-reported mood. *Journal of Personality and Social Psychology*, *54*(2), 296–308. https://doi.org/10.1037/0022-3514.54.2.296

Daňsová, P., Masopustová, Z., Hanáčková, V., Kicková, K., & Korábová, I. (2016). Metoda patient health questionnaire-9: Česká verze [The patient health questionnaire-9: The Czech version]. *Ceskoslovenska Psychologie*, *60*(5), 468–481.

Elliott, R. (2002). Hermeneutic single-case efficacy design. *Psychotherapy Research*, *12*(1), 1–21. https://doi.org/10.1080/713869614

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardized brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, *180*(1), 51–60. https://doi.org/10.1192/bjp.180.1.51

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Flückiger, C., Regli, D., Grawe, K., & Lutz, W. (2007). Similarities and differences between retrospective and pre-post measurements of outcome. *Psychotherapy Research*, *17*(3), 359–364. https://doi.org/10.1080/10503300600830728

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520–531. https://doi.org/10.1037/a0031669

Grawe, K. (2004). *Psychological therapy*. Hogrefe & Huber.

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice*, *10*(2), 33–41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Howard, G. S., Millham, J., Slaten, S., & O'Donnell, L. (1981). Influence of subject response style effects on retrospective measures. *Applied Psychological Measurement*, *5*(1), 89–100. https://doi.org/10.1177/014662168100500113

Hsu, L. M. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy. *Journal of Consulting and Clinical Psychology*, *63*(1), 141–144. https://doi.org/10.1037/0022-006X.63.1.141

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59* (1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling*. R package version 0.5-6. https://CRAN.R-project.org/package = semTools.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kiresuk, T., & Lund, S. (1979). Goal attainment scaling: Research, evaluation and utilization. In H. C. Schulberg, & F. Parker (Eds.), *Program evaluation in health fields (Vol. 2* (pp. 214–237). Human Science Press.

Kiresuk, T., & Sherman, R. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, *4* (6), 443–453. https://doi.org/10.1007/BF01530764

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*(1), 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Krampen, G. (2010a). Direct measurement of psychotherapeutic outcomes: Experimental construction and validation of a brief scale. *Psychological Test and Assessment Modeling*, *52*(1), 29–47.

Krampen, G. (2010b). Experimentelle Konstruktion eines Kurzfragebogens zur direkten Veränderungsmessung psychotherapeutischer Effekte im Befinden. *Diagnostica*, *56*(4), 212–221. https://doi.org/10.1026/0012-1924/a000024

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kusier, A. O., & Folker, A. P. (2020). The well-being index WHO-5: Hedonistic foundation and practical limitations. *Medical Humanities*, *46*(3), 333–339. https://doi.org/10.1136/medhum-2018-011636

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 and Psychotherapy*, *3*(4), 249–258. https://doi.org/10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care*, *46*(3), 266–274. https://www.jstor.org/stable/40221654

Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, *9*(2), 111–131. https://doi.org/10.1037/1089-2680.9.2.111

Masroor, U., Khan, M. J., & Iqbal, N. (2013). Psychotherapeutic intervention outcomes among clinical patients with and without personality disorder co-morbidity. *Pakistan Journal of Clinical Psychology*, *12*(2), 43–53.

McCullough, G., Huebner, E. S., & Laughlin, J. E. (2000). Life events, self-concept, and adolscents positive subjective well-being. *Psychology in the Schools*, *37*(3), 281–290. https://doi.org/10.1002/(SICI)1520-6807(200005)37:3<281::AID-PITS8>3.0.CO;2-2

Michalak, J., Kosfelder, J., Meyer, F., & Schulte, D. (2003). Messung des Therapieerfolgs. Veränderungsmasse oder retrospektive Erfolgsbeurteilung [Measurement of therapy success: Pre-post effect sizes and retrospective success evaluation]. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *32*(2), 94–103. https://doi.org/10.1026/0084-5345.32.2.94

Miller, S. D., Duncan, B. L., Brown, J., Sparks, J. A., & Claud, D. A. (2003). The Outcome Rating Scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy*, *2*(2), 91–100.

Nesselroade, J. R., & Ghisletta, P. (2003). Structuring and measuring change over the life span. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development* (pp. 317–337). Springer. https://doi.org/10.1007/978-1-4615-0357-6_14

Neusar, A. (2014). To trust or not to trust? Interpretations in qualitative research. *Human Affairs*, *24*(2), 178–188. https://doi.org/10.2478/s13374-014-0218-9

Pourová, M., Řiháček, T., Boehnke, J. R., Šimek, J., Saic, M., Kabát, J., & Šilhán, P. (2024). Effectiveness of a multicomponent group-based treatment in patients with medically unexplained physical symptoms: A multisite naturalistic study. *Journal of Contemporary Psychotherapy*, *54*, 47–57. https://doi.org/10.1007/s10879-023-09597-4

Prikner, O. (2021). *Vybrané psychometrické charakteristiky škály GAD-7* [Selected psychometric characteristics of the GAD-7 scale; Master's thesis]. Masaryk University. https://is.muni.cz/th/imwek/Diplomova-prace-finis.pdf.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Řiháček, T., Čevelíček, M., Boehnke, J. R., Pourová, M., & Roubal, J. (2022). Mechanisms of change in multicomponent group-based treatment for patients suffering from medically unexplained physical symptoms. *Psychotherapy Research*, *32* (8), 1016–1033. https://doi.org/10.1080/10503307.2022.2061874

Rodgers, B., & Elliott, R. (2015). Qualitative methods in psychotherapy outcome research. In O. C. G. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy research: Foundations, process, and outcome* (pp. 559–578). Springer.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Roubal, J., Řiháček, T., Čevelíček, M., Hytych, R., & Holub, D. (2018). Retrospective client interviewing can inform clinicians' practice and complement routine outcome monitoring. *Revista Argentina de Clínica Psicológica*, *27*(2), 294–320.

Safer, M. A., & Keuler, D. J. (2002). Individual differences in misremembering pre-psychotherapy distress: Personality and memory distortion. *Emotion*, *2*(2), 162–178. https://doi.org/10.1037/1528-3542.2.2.162

Sandell, R., & Wilczek, A. (2016). Another way to think about psychological change: Experiential vs. incremental. *European Journal of Psychotherapy & Counselling*, *18*(3), 228–251. https://doi.org/10.1080/13642537.2016.1214163

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Stänicke, E., & McLeod, J. (2021). Paradoxical outcomes in psychotherapy: Theoretical perspectives, research agenda and practice implications. *European Journal of Psychotherapy & Counselling*, *23*(2), 115–138. https://doi.org/10.1080/13642537.2021.1923050

Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 Well-Being Index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, *84*(3), 167–176. https://doi.org/10.1159/000376585

Vanier, A., Oort, F. J., McClimans, L., Ow, N., Gulek, B. G., Böhnke, J. R., Sprangers, M., Sébille, V., Mayo, N., & the Response Shift - in Sync Working Group (2021). Response shift in patient-reported outcomes: Definition, theory, and a revised model. *Quality of Life Research*, *30*(12), 3309–3322. https://doi.org/10.1007/s11136-021-02846-w

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement*, *37*, 679–689. https://doi.org/10.1177/001316447703700310

Willutzki, U., Ülsmann, D., Schulte, D., & Veith, A. (2013). Direkte Veränderungsmessung in der Psychotherapie: Der Bochumer Veränderungsbogen-2000 (BVB-2000). *Zeitschrift für Klinische Psychologie und Psychotherapie*, *42*(4), 256–268. https://doi.org/10.1026/1616-3443/a000224

Zielke, M., & Kopf-Mehnert, C. (1978). *Veränderungsfragebogen des Erlebens und Verhaltens (VEV) [Questionnaire to assess changes in experiencing and behavior (QCEB)]*. Beltz.