Československá psychologie 2021 / ročník LXV / číslo 4
DOI: 10.51561/cspsych.65.4.353

CZECH VERSION OF THE OUTCOME RATING SCALE: SELECTED PSYCHOMETRIC PROPERTIES

DANA SERYJOVÁ JUHOVÁ[1], TOMÁŠ ŘIHÁČEK[1], HYNEK CÍGLER[1], EVA DUBOVSKÁ[2], MARTIN SAIC[3], MARTIN ČERNÝ[4,5,6], JAN DUFEK[7], SCOTT D. MILLER[8]

[1]*Department of Psychology, Faculty of Social Studies, Masaryk University, Brno, Czech Republic*
[2]*Prague College of Psychosocial Studies, Prague, Czech Republic*
[3]*Daily Sanatorium Horní Palata, General University Hospital in Prague, Prague, Czech Republic*
[4]*Department of Psychiatry, First Faculty of Medicine, Charles University in Prague, Czech Republic*
[5]*General University Hospital in Prague, Czech Republic*
[6]*Department of Psychiatry, Teaching Hospital Královské Vinohrady, Prague, Czech Republic*
[7]*Psychiatric Clinic, University Hospital Brno, Czech Republic*
[8]*International Center for Clinical Excellence, Chicago, Illinois, USA*

ABSTRACT

Objectives. The Outcome Rating Scale (ORS) is an ultra-brief self-report scale designed to measure change during psychotherapy. The goal of this study was to test (a) the factor structure of the ORS, (b) the measurement invariance between a clinical and a non-clinical sample, between pre-therapy and post-therapy assessment (within the clinical sample), and between online and paper-and-pencil forms of administration (within the non-clinical sample), (c) concurrent validity with other outcome measures, and (d) sensitivity to therapeutic change.
Sample and settings. $N = 256$ patients, $N = 210$ non-clinical respondents, and $N = 89$ students participated in the study. Patients responded to the ORS before and after psychotherapy.
Statistical analysis. The factor structure and measurement invariance were tested using confirmatory factor analysis. Concurrent validity and test-retest reliability were assessed using correlational analysis. Sensitivity to change was assessed using the Reliable Change Index and pre-post effect size.
Results. The unidimensional structure was supported. The best-fitting model was a partially tau-equivalent model with the first and the fourth items' loadings fixed to the same value. While only metric invariance was demonstrated between the clinical and non-clinical samples, the ORS demonstrated scalar invariance between pre- and post-therapy assessment and strict invariance between the paper-and-pencil and online forms of administration. Internal consistency, as well as concurrent validity, were satisfactory. The sensitivity to the therapeutic change was adequate. Furthermore, internal consistency and sensitivity to change were increased if the score was computed as a weighted sum of items.
Study limitation. The samples were not representative.

key words:
confirmatory factor analysis,
measurement invariance,
Outcome Rating Scale,
ORS

klíčová slova:
konfirmační faktorová analýza,
invariance měření,
Outcome Rating Scale,
ORS

Psychotherapy has entered the era of routine outcome measurement and feedback-informed treatment (e.g., Lambert, 2015; Prescott et al., 2017; Scott & Lewin, 2015). Psychotherapists are expected to integrate outcome measurement routines in their clinical practice and, if possible, to utilize this data to improve psychotherapy out-

Submitted: 17. 10. 2020; D. S. J., Department of Psychology, Faculty of Social Studies, Masaryk University, Joštova 10, 602 00 Brno; e-mail: juhova@fss.muni.cz

This study was funded by The Czech Science Foundation (grant GA18-08512S).
We thank our colleagues who helped us with the data collection: Věra Žežulková, Markéta Elsnerová, Lenka Juhová, Marie Malinová, Tomáš Peřich, Hana Provazníková, Petr Šíftek, and Martin Kuška.

comes. To achieve this goal, several brief outcome measures were developed, including electronic administration systems that allow psychotherapists to obtain instant feedback on the progress of a case (see Lyon et al., 2016, for a review). Measures most often used to track patient progress include Outcome Questionnaire-45 (OQ-45; Lambert et al., 1996), Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM; Barkham et al., 2001), Treatment Outcome Package (TOP; Kraus et al., 2005), and Outcome Rating Scale (ORS; Miller et al., 2003). This study focused on the validation of the Czech version of the ORS, the shortest and most practice-friendly one of these measures.

The ORS is a four-item self-report scale that measures a client's overall psychological well-being (Miller et al., 2003). Conceptually, it follows the dimensions of the OQ-45, which include subjective discomfort, interpersonal relations, and social role performance. Given its brevity, the ORS is particularly suited to track clients' progress on a session-by-session basis.

Psychometric studies have demonstrated acceptable psychometric properties of the scale, representing a balanced trade-off between the reliability and validity of longer outcome measures and the feasibility of an ultra-brief scale. The ORS demonstrated an excellent internal consistency in a non-clinical sample (Cronbach's $\alpha > .90$), satisfactory test-retest reliability after one week ($r_{tt'}$ between .66 and .80), good concurrent validity with other outcome measures (e.g., $r \geq -.59$ with OQ-45), and an ability to distinguish between clinical and non-clinical samples (Bringhurst et al., 2006; Miller et al., 2003). These results were essentially replicated in psychometric studies using other language versions (Biescad & Timulak, 2014; Janse et al., 2014). The Slovak version of the scale was found to be sensitive to change, yielding large effect sizes (Biescad & Timulak, 2014). Furthermore, it was strongly associated with the OQ-45 (-.69) and CORE-OM (-.70) scores in a clinical sample (Bieščad, 2007).

Traditionally, the measurement of the therapeutic change is based on the assumption that the scale, the ORS in this case, measures the same construct across samples and conditions (cf. Brown, 2015). However, this assumption has been questioned. Sandell and Wilczek (2016) argued that clients change *qualitatively* in psychotherapy, and they answer outcome measures from different perspectives, before and after psychotherapy. Therefore, the incremental change in outcome scores may not properly reflect the change as experienced by a client, which may explain why some studies have reported discrepancies between quantitative and qualitative assessments of therapeutic changes (e.g., Doran et al., 2015; Hill et al., 2013). Therefore, it is essential to assess the factorial invariance of outcome measures across measurement conditions, such as pre-therapy and post-therapy.

Furthermore, the concepts of statistically reliable change and clinically significant change (Jacobson & Truax, 1991) are often used to evaluate the change status of a client. Again, these concepts assume factorial invariance. The most widely used criterion of clinically significant change (criterion "*c*" in Jacobson and Truax's nomenclature) is derived from the score distributions of both the clinical and non-clinical populations, which is meaningful only if the scale measures the same construct across these populations. Similarly, the calculation of the Reliable Change Index (RCI), as originally formulated by Jacobson and Truax, is based on the test-retest reliability of a measure, which is often derived from non-clinical samples and extrapolated to patients (although this problem may be overcomed by using internal consistency estimates obtained from a clinical sample as a reliability index, Schauenburg & Strack, 1999).

To date, no study has tested the factor structure of the ORS and its invariance across samples and measurement conditions. Because the ORS was designed as a one-di-

mensional scale (Miller et al., 2003), we expect the measure to have a unidimensional structure. However, item loadings, intercepts, and residual variance may differ across groups and conditions. While invariance testing is still not a part of standard psychometric evaluations of outcome measures, its use has been increasing. For instance, it has been used to compare the factor structure of some outcome measures between clinical and non-clinical samples (e.g., Rice et al., 2014), across age and gender (e.g., O'Reilly et al., 2016), as well as time (e.g., Jabrayilov et al., 2017). However, such an evaluation of the ORS is still missing.

The primary aim of the study was to test the hypothesized unidimensional factor structure of the Czech version of the ORS and its invariance between the clinical and the non-clinical sample, between two measurement points (pre-therapy vs. post-therapy), and between two forms of administration (paper-and-pencil vs. online). Since the results suggested that it might be more appropriate to use a weighted sum score instead of a raw sum of items, we also explored the consequences of this choice on various psychometric properties, including validity, reliability, and sensitivity to change.

METHOD

Participants

The study included three independent samples. Sample 1 (hereafter referred to as the clinical sample) consisted of adult group psychotherapy patients recruited at four psychotherapy clinics. Out of 541 patients, 168 (31.1%) patients refused to participate, and 22 (4.1%) patients were excluded due to missing data. Furthermore, patients with a psychotic disorder ($n = 74$, 13.7%) and those for whom a diagnosis was not provided ($n = 21$,3.9%) were excluded to maintain sample homogeneity. This resulted in a clinical sample of $N = 256$ patients. Out of this number, $n = 43$ (16.8%) were inpatients and $n = 213$ (83.2%) were outpatients receiving intensive psychotherapy treatment on a daily basis. Out of the clinical sample, $n = 168$ (65.6%) patients also completed the measurement battery at treatment termination.

Sample 2 (hereafter referred to as the non-clinical sample) consisted of adults who, based on their self-reports, had never received a psychiatric diagnosis, did not suffer from any serious mental problems, and had not been in psychotherapy during the previous 12 months. Out of 290 participants, 80 (27.6%) were excluded due to missing data or because they did not meet the abovementioned requirements. This resulted in a non-clinical sample of $N = 210$. Out of this number, $n = 119$ (56.7%) completed paper-and-pencil questionnaires, while $n = 91$ (43.3%) completed the questionnaires online.

Sample 3 (hereafter referred to as the student sample) was created solely to establish the test-retest reliability. This sample consisted of undergraduate students in psychology and pedagogy. To be included, they had to meet the conditions for the non-clinical sample and be between 18 and 26 years old. Furthermore, they had to participate in at least two repeated measurements. Of the 91 students who responded to the survey, $N = 89$ students provided two measurements ($n = 62$ completed the questionnaire online, $n = 29$ completed the paper-and-pencil form) and $N = 79$ provided three measurements. See Table 1 for additional information on all three samples and Figure 1 for a flow diagram.

Procedure

The clinical sample was recruited at four psychotherapy clinics. Patients who agreed to participate completed the questionnaire battery in a paper-and-pencil form before the beginning of the treatment and at treatment termination.

Table 1 Characteristics of the clinical and non-clinical samples

Socio-demographic Characteristics	Category	Clinical sample *N* = 256	Non-clinical sample *N* = 210	Student sample[a] *N = 91*
Sex[b]	Female	179 (69.9%)	153 (72.9%)	79 (86.8%)
	Male	77 (30.1%)	56 (26.7%)	12 (13.2%)
Age	Range	18 – 65	18 – 72	18 – 26
	M (SD)	38.0 (12.1)	36.5 (11.7)	21.5 (1.9)
Diagnostic category	F0	4 (1.6%)	–	–
	F1	1 (0.4%)	–	–
	F3/F4	189 (73.8%)	–	–
	F5	16 (6.2%)	–	–
	F6	46 (18.0%)	–	–

Note. F0 = organic, including symptomatic, mental disorders; F1 = mental and behavioral disorders due to psychoactive substance use; F3 = mood disorders; F4 = neurotic, stress-related and somatoform disorders (F3 and F4 were merged because some clinics did not differentiate between the two categories); F5 = behavioral syndromes associated with physiological disturbances and physical factors; F6 = disorders of adult personality and behavior.
[a] First measurement.
[b] One respondent in the non-clinical sample did not report sex.

The non-clinical sample was recruited in a snowball manner using our personal contacts, as well as social media. In this sample, we combined paper-and-pencil and online data collection. The intention was to allow respondents to choose the form they were most comfortable with and the decision to test measurement invariance between both forms was made later on. The combined form and the snowball character of the data collection did not allow us to report the response rate.

The student sample was recruited at two universities. Student participation was voluntary, without any consequences following from a refusal to participate. Students were asked to fill out the questionnaire three times in weekly intervals.

Measures

Outcome Rating Scale. The ORS is a self-report outcome measure composed of four visual analogue scales. The items focus on the individual, relational, social, and overall well-being of a client. Each item is represented by a 10-centimeter horizontal line, without any verbal anchors. The total score is computed as the sum of all items, ranging from 0 to 40 (or 0 to 400 in case the response is measured in millimeters). Respondents are instructed to report their well-being "looking back over the last week" (Miller et al., 2003). The Czech version of the ORS (Zatloukal et al., 2006) was used in this study. The translation was conducted in collaboration with one of the ORS authors and followed the rules of the International Center for Clinical Excellence (ICCE).

Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM). CORE-OM (Barkham et al., 2001) contains 34 items divided into four domains: well-being, problems/symptoms, functioning, and risk. All items are rated using a five-point Likert-type scale ranging from "not at all" to "most or all the time". The total score is computed as the average of all items, ranging from 0 to 4. Higher scores

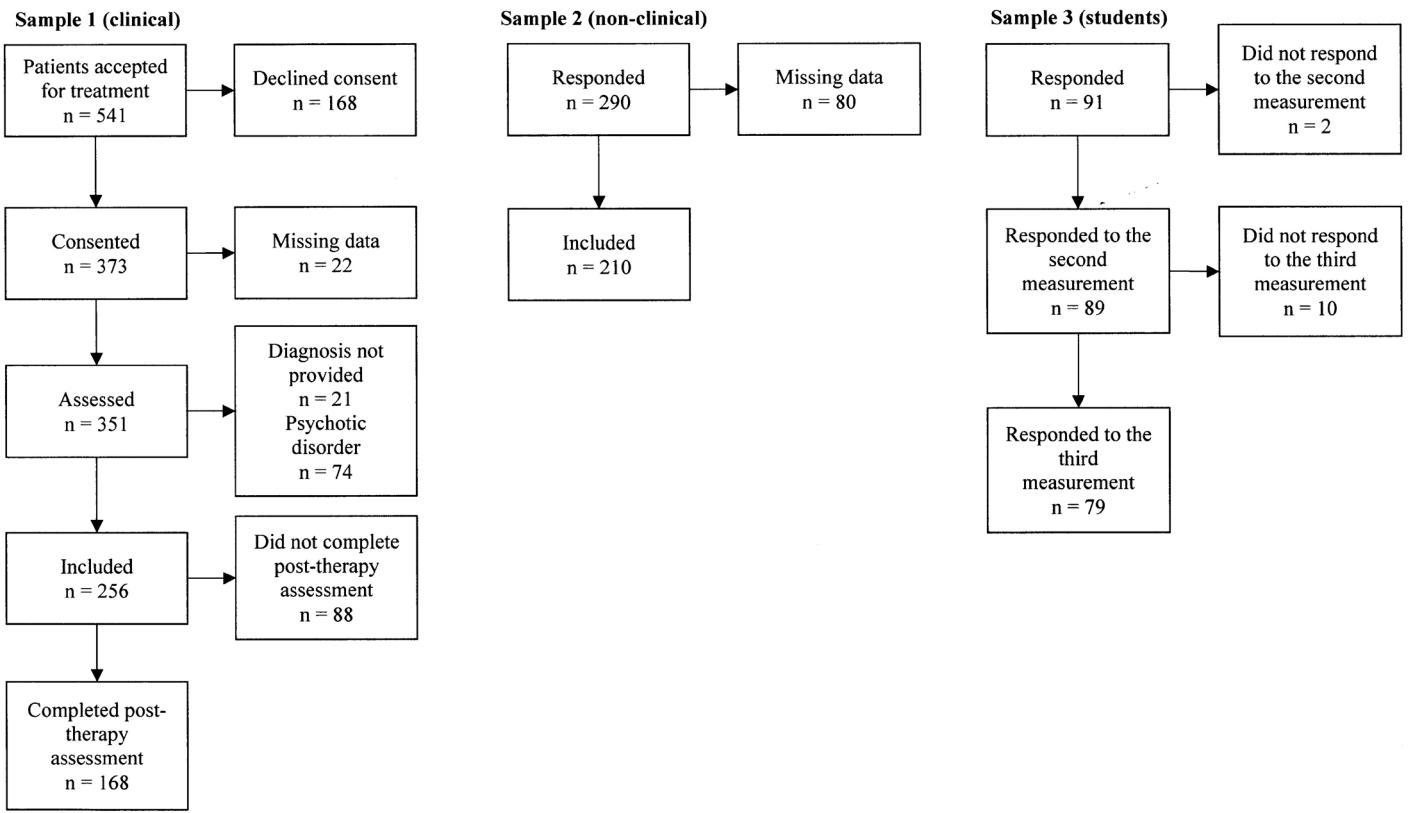

Figure 1 Sample flow diagram

represent a higher rate of pathology. The theoretical structure was not confirmed empirically, and the use of subscale scores is not recommended (Juhová et al., 2018). In this study, we used the total score without the Risk subscale (6 items) and internal consistency of the scale was $\alpha = .925$ ($\omega_{tot} = .935$) for the clinical sample and $\alpha = .875$ ($\omega_{tot} = .901$) for the non-clinical sample.

Rosenberg Self-Esteem Scale (RSES). RSES (Rosenberg, 1965) is a Likert-type scale used for mapping global self-esteem. RSES contains 10 items, five negative formulations and five positive. The response options for each question include a four-point scale ranging from "strongly agree" to "strongly disagree". Although there is a debate concerning the number of dimensions, existing research favors a unidimensional structure (Sinclair et al., 2010). The total score was calculated as the average of all items, ranging from 0 to 3. The higher the score, the higher the self-esteem. Internal consistency of the total score was $\alpha = .846$ ($\omega_{tot} = .850$) for the clinical sample and $\alpha = .822$ ($\omega_{tot} = .833$) for the non-clinical sample.

Symptom-Checklist-90 (SCL-90). SCL-90 is a 90-item scale designed to map the level of nine primary symptom dimensions, as experienced during the last seven days (Derogatis & Cleary, 1977). However, the subscale scores are not independent and the scale is recommended exclusively for overall distress assessment (Groth-Marnat, 2009). Respondents indicate the degree to which they experience each of the symptoms on a scale ranging from "not at all" (0) to "extremely" (4). The higher the score, the higher the rate of psychopathology. In this study, we used the total score (usually called Global Severity Index, GSI) calculated as the average of all items. GSI internal consistency was $\alpha = .966$ and McDonald omega total $\omega_{tot} = .970$ for the clinical sample.

Demographic questionnaire. For the purpose of this study, we created a 10-item questionnaire to map basic demographic data. The non-clinical version contained additional questions that assessed the state of a respondent's mental health.

Patients in the clinical sample were administered all measures. Respondents in the non-clinical sample were administered all measures except for the SCL-90. Respondents in the student sample answered the ORS only.

Statistical analysis

We analyzed data from only those respondents who filled out the whole ORS and more than 90% of items for each of the other measures. Sample sizes mentioned in the Participants section include only those respondents who met these requirements. For respondents who answered more than 90% but less than 100% of items, total scores we computed as an average of all answered items.

We tested the ORS factor structure using confirmatory factor analysis (CFA). Since the ORS items are highly correlated and cannot be treated as separate dimensions (Miller et al., 2003), we tested multiple versions (congeneric, tau-equivalent, and partially tau-equivalent) of a one-factor model. After selecting the model with the best fit, we tested measurement invariance: (a) between the clinical and non-clinical samples, (b) between two measurement points (pre-therapy vs. post-therapy) within the clinical sample, and (c) between two modes of administration (paper-and-pencil vs. online) within the non-clinical sample. We gradually put equality constraints on the following parameters: factor loadings (metric invariance), item intercepts (scalar invariance), residual variances (strict invariance), and latent means. We used the robust maximum likelihood estimation (MLR).

Model fit was assessed using χ^2 statistics, the Tucker-Lewis-Index (TLI), the Standardized Root Mean Square Residual (SRMR), and the Bayesian Information Criterion

(BIC). A good fit was indicated by values greater than or equal to .95 for TLI, and less than or equal to .08 for SRMR (Hu & Bentler, 1998). Smaller BIC values indicate a better fit. We did not report RMSEA because it is a problematic indicator in models that have limited degrees of freedom (Kenny et al., 2015).

Measurement invariance was assessed by the change in fit compared to a previous model: a change in TLI ≥ .010 for all levels of invariance tests, a change in SRMR ≥ .030 for metric invariance, and ≥ .010 for scalar and strict invariance indicate non-invariance between groups (Chen, 2007). We used robust estimates using the Satorra -Bentler method (Satorra & Bentler, 2001). To estimate the similarity of factor solutions between the clinical and non-clinical samples, we used the Tucker coefficient of congruence (r_c, see, e.g., Lorenzo-Seva & ten Berge, 2006).

After discovering that ORS invariance was unsatisfactory between the clinical and non-clinical samples (see Table 2), we performed a latent regression analysis to estimate the respective contribution of each ORS item to the latent score for both samples separately. We then used non-standardized regression coefficients as item weights to compute a weighted ORS score. To facilitate interpretation, we adjusted the weights to produce scores in the range of 0 to 40, which corresponds to the range of raw sum scores.

Next, we explored the consequences of using the weighted score for various psychometric properties of the ORS. First, to assess concurrent validity, we computed Pearson correlation coefficients between the ORS scores (weighted and raw) and CORE-OM, RSES, and SCL-90 scores. Second, we assessed associations of the ORS scores with gender and age. Third, we assessed the internal consistency of the weighted score using Raykov's omega and compared it to Cronbach's alpha of the raw score. To assess test-retest reliability, we computed Pearson correlation coefficients for pairs of repeated measurements in the student sample. Fourth, to assess sensitivity to change, we computed the effect size (Cohen's *d*) of change after psychotherapy in the clinical sample. Fifth, we computed the RCI and the clinical cutoff according to Jacobson and Truax (1991):

$$RCI = 1.96\sqrt{2SD^2(1 - rel)} \tag{1}$$

where *SD* = standard deviation of the pre-treatment scores in the clinical sample, *rel* = reliability of the measure, and 1.96 = the corresponding quantile of normal distribution (in this case 95%). Jacobson and Truax (1991) recommended to use the test-retest correlation obtained on a non-clinical sample as the estimate of reliability. However, the measurement invariance between the clinical and non-clinical samples was not satisfactory in our case and, therefore, the use of a parameter obtained on the non-clinical sample was not warranted. Following Evans et al.'s (1998) and Schauenburg & Strack's (1999) recommendation, we used internal consistency coefficients instead.

To estimate the clinical cutoff score, we used the criterion "c" proposed by Jacobson and Truax (1991) for cases in which the distributions of the clinical and non-clinical populations overlap. We proceeded with the following formula:

$$c = \frac{SD_{non\text{-}clinical}M_{clinical} + SD_{clinical}M_{non\text{-}clinical}}{SD_{non\text{-}clinical} + SD_{clinical}} \tag{2}$$

The analysis was conducted using R (version 3.4.2), with the lavaan (Rosseel, 2012) and lm.beta (Behrendt, 2014) packages.

RESULTS

Confirmatory factor analysis

Using CFA, we supported the unidimensional model of the ORS on both samples separately. We tested three models: a congeneric, a tau-equivalent, and a partially tau-equivalent model. The congeneric model had a negative residual variance on Item 4 (overall well-being) in the clinical sample; thus, the model could not be interpreted adequately. The tau-equivalent model, in which all loadings had the same unstandardized values, did not fit the data because Item 4 (overall well-being) had a substantially higher loading and a substantially lower or negative residual variance compared to the remaining items. Finally, we tested a partially tau-equivalent model. The only difference from the congeneric model was the constrained loading of Item 1 (personal well-being) and Item 4 (overall well-being) to the same value. This solved the problem with negative variance in the congeneric model, and the model still had an acceptable fit. For this reason, we used this model in the following analyses. The fit indices of all analyses are provided in Table 2.

Table 2 Fit indices for CFA models of the ORS on the clinical and non-clinical samples

Model	Sample	χ^2	*df*	TLI	SRMR
Congeneric	Clinical[ac]	2.13	2	.999	.019
	Non-clinical[b]	13.24***	2	.912	.035
Tau-equivalent	Clinical[a]	48.80***	5	.890	.159
	Non-clinical[b]	33.29***	5	.923	.159
Partially tau-equivalent[d]	Clinical[a]	11.58**	3	.959	.058
	Non-clinical[b]	15.14**	3	.938	.040

Note. [a] $N = 256$.
[b] $N = 210$.
[c] The congeneric cannot be interpreted adequately because of a negative residual variance on the fourth item.
[d] The partially tau-equivalent model had Items 1 and 4 fixed to the same value of factor loading.
**$p < .01$, ***$p < .001$.

Invariance between the clinical and the non-clinical sample

Next, we tested the invariance between the clinical and non-clinical samples, based on the partially tau-equivalent model. Table 3 shows the fit statistics.

TLI and SRMR values, as well as the change in TLI, were acceptable for metric and scalar invariance. However, the change in SRMR indicated poor fit for scalar invariance, meaning that the item intercepts differed across the samples. We therefore concluded that only metric invariance was demonstrated between the samples. Figure 2 illustrates the metric invariance between the samples (there was minimal overlap between the samples).

The similarity of factor solutions for the clinical and non-clinical samples was high (congruence $r_c = .98$). The means of the latent attribute in the clinical and non-clinical samples differed: in the scalar invariant model, Cohen's $d = 1.85$, 95% CI [1.55, 2.16] (approximately 39 millimeters). Although the measurement is not scalar invariant and, thus, the average difference between the clinical and the nonclinical sample is biased, the difference is too large to be attributed to noninvariance alone. Moreover, the degree of noninvariance is relatively small, given that only the SRMR value was poor.

Table 3 Fit indices for testing invariance between the clinical and non-clinical samples

Invariance	χ^2	*df*	$\Delta\chi^2$	Δdf	BIC	TLI	ΔTLI	SRMR	ΔSRMR
Invariance between the clinical and non-clinical sample									
Configural	27.32	6			16342	.947		.050	
Metric	41.36	8	15.58***	2	16344	.942	.005	.082	.032
Scalar	51.38	11	9.69**	3	16336	.950	.008	.087	.005
Strict	94.13	15	36.39***	4	16377	.918	.032	.097	.009
Means	301.97	16	434.76***	1	16593	.731	.187	.705	.609
Invariance between pre- and post-therapy measurement in the clinical sample									
Configural	62.15	18			11820	.917		.084	
Metric	63.01	20	0.28	2	11810	.928	.011	.085	.001
Scalar[a]	65.28	23	2.18	3	11797	.939	.011	.081	.004
Strict	327.77	27	80.19	4	12346	.312	.627	.286	.205
Invariance between paper-and-pencil and electronic form in the non-clinical sample									
Configural	16.29	6			7191	.945		.041	
Metric	18.32	8	1.35	2	7182	.961	.016	.050	.009
Scalar	23.37	11	4.50	3	7171	.969	.008	.052	.003
Strict	24.56	15	2.59	4	7154	.981	.012	.048	-.004
Means	29.50	16	6.24*	1	7155	.975	-.006	.089	.041

Note. To test invariance, the partially tau-equivalent model was used.
[a] The model with the fixed latent means had negative residual item variances.
*p < .05, **p < .01, ***p < .001.

To estimate the respective contribution (i.e., weight) of each ORS item to the latent score, we performed a latent regression for both samples separately. The weights, linearly transformed to preserve the range of scores from 0 to 40, were as follows: Item 1 = 0.16, Item 2 = 0.04, Item 3 = 0.04, and Item 4 = 3.76 for the clinical sample and Item 1 = 1.25, Item 2 = 0.42, Item 3 = 0.21, and Item 4 = 2.13 for the non-clinical sample.

Invariance across time

Within the clinical sample, we tested the invariance between the pre-therapy (n = 256) and post-therapy measurements (n = 168). First, we estimated the fit of the partially tau-equivalent factor model for the post-therapy data. The model fit was satisfactory, $\chi^2(3) = 18.36$, $p < .001$, TLI = .917, SRMR = .085.

To test invariance across time, we designed a simple structural model with latent regression in which the first factor loaded on items from the pre-therapy measurement, the second factor loaded on items from the post-therapy measurement, and, at the same time, the first factor predicted the second factor. Allowing all residual correlations led to a predicted correlation matrix that was not positively defined. Therefore, we fixed residual correlations for Item 4 (which had the highest factor loadings) to

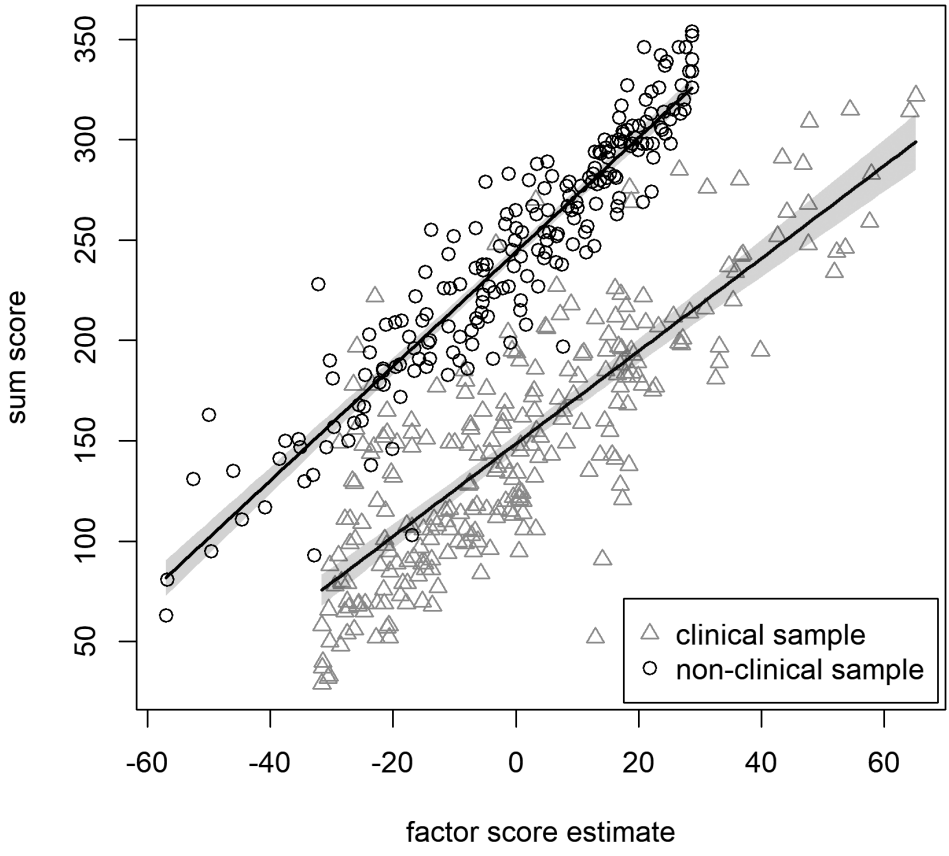

Figure 2 Correlation between item sums and factor score estimates for both samples

zero in both measurement models, which eliminated this problem. This model described the data better than the model without residual correlations, $\Delta\chi^2(3) = 31.0$, $p < .001$. Table 3 shows the fit statistics. None of the steps led to a significant deterioration of model fit.

The results show that while the ORS measures the same attribute before and after therapy, the model is not strictly invariant (i.e., the precision of measurement differs at the beginning and the end of therapy). The error variances in the scalar model were smaller at the end of the therapy and, together with higher factor variance ($SD_{time1} = 22.05$, $SD_{time2} = 25.90$), this led to higher reliability of the post-therapy measurement.

Invariance across forms of administration

Within the non-clinical sample, in which both the paper-and-pencil and online forms of administration were used, we tested the measurement invariance between these two conditions. The results show that the measure performed similarly across the two forms of administration. Table 3 shows the fit statistics. We can conclude that both versions measured the same trait with the same precision and that the validity of the remaining analyses was not threatened by pooling the data obtained through these

two forms of administration. Interestingly, latent means slightly differed; the participants who completed the questionnaire online had approximately an 8-millimeter (d = -0.33, 95% CI [-0.57, -0.08]) lower average score in the strict model.

Consequences of using raw vs. weighted total score

In this section, we explored the consequences of using raw versus weighted scores for various psychometric characteristics of the ORS. Descriptive statistics of all samples and measures used in this section are presented in Table 4.

Table 4 Descriptive statistics of the initial measurement

Sample	Measure	*M*	*SD*	*Min*	*Max*	Skewness	Kurtosis
Clinical (N = 256)	ORS (weighted sum)	127.79	88.72	0.00	391.11	0.71	0.00
	ORS (raw sum)	142.37	80.66	0.00	379.00	0.59	-0.05
	CORE-OM	2.19	0.71	0.18	6.11	0.12	3.02
	RSES	1.42	0.53	0.10	2.80	0.11	-0.14
	SCL-90 GSI	1.47	0.65	0.22	4.71	0.75	1.86
Non-clinical (N = 210)	ORS (weighted sum)	280.66	84.12	44.00	400.00	-0.68	-0.12
	ORS (raw sum)	278.94	79.66	47.00	400.00	-0.54	-0.36
Student (N = 91)	ORS (weighted sum)	280.98	67.07	94.63	400.00	-0.43	-0.16
	ORS (raw sum)	281.59	63.11	95.00	400.00	-0.33	-0.05

Concurrent validity. We explored the association of the ORS scores with three reference measures, CORE-OM, RSES, and, for the clinical sample, SCL-90. The results are presented in Table 5. The associations were strong, in the expected direction, and essentially the same for both versions of the ORS score.

Table 5 Concurrent validity of the ORS score with reference instruments

Sample	ORS total score	*r* [95% CI]		
		CORE-OM [a]	RSES	SCL-90
Clinical (N = 256)	Weighted sum	-.68 [-.74, -.61]	.47 [.36, .56]	-.49 [-.58, -.39]
	Raw sum	-.67 [-.73, -.60]	.44 [.33, .53]	-.47 [-.56, -.37]
Non-clinical (N = 210)	Weighted sum	-.54 [-.63, -.43]	.49 [.38, .58]	–
	Raw sum	-.54 [-.63, -.44]	.48 [.36, .57]	–

Note. All coefficients are significant at $p < .001$.
[a] Total score without Risk items.

Associations with gender and age. The differences between men and women were non-significant, with small to zero effect sizes in both samples: d = -0.02, 95% CI [-0.29, 0.25] for the weighted score and d = 0.00, 95% CI [-0.26, 0.27] for the raw

score in the clinical sample; $d = 0.30$, 95% CI [-0.01, 0.61] for the weighted score and $d = 0.29$, 95% CI [-0.02, 0.60] for the raw score in the non-clinical sample (women had higher scores in the non-clinical sample).

Similarly, there was no significant relationship between age and either version of the ORS score. In the clinical sample, the correlations were $r = -.02$, 95% CI [-.14, .11], $p = .79$ for the weighted score and $r = -.03$, 95% CI [-.15, .09], $p = .62$ for the raw score. In the non-clinical sample, it was $r = .06$, 95% CI [-.08, .20], $p = .38$ for the weighted score and $r = .06$, 95% CI [-.07, .20], $p = .37$ for the raw score.

Reliability. To assess the internal consistency of the ORS, we compared Raykov's omega of the latent (i.e., weighted) score to Cronbach's alpha of the raw score. In the clinical sample, $\omega = .98$ and $\alpha = .78$. In the non-clinical sample, $\omega = .94$ and $\alpha = .88$. The weighted score was superior in terms of internal consistency.

To assess the stability of the ORS scores in time, we used data from the student sample. The coefficients of test-retest reliability after one week (i.e., between the first and the second administration) were $r_{tt'} = .56$, 95% CI [.40, .69] for the weighted score and $r_{tt'} = .61$, 95% CI [.46, .73] for the raw score. After two weeks (i.e., between the first and the third administration), they were $r_{tt'} = .61$, 95% CI [.45, .73] for the weighted score and $r_{tt'} = .66$, 95% CI [.51, .77] for the raw score. Here, the raw score slightly outperformed the weighted score but, given the overlap of the confidence intervals, no definite conclusions can be drawn.

Sensitivity to change. To assess sensitivity to change, we compared pre- and post-treatment scores in the clinical sample. The effect sizes were as follows: $d = -0.88$, 95% CI [-1.11, -0.66] for ORS (weighted score), $d = -0.91$, 95% CI [-1.14, -0.68] for ORS (raw score), $d = 0.72$, 95% CI [0.50, 0.94] for CORE-OM, $d = -0.45$, 95% CI [-0.67, -0.24] for RSES, and $d = 0.85$, 95% CI [0.63, 1.08] for SCL-90 GSI. While both versions of the ORS score appear to be superior in terms of the effect size compared to the other outcome measures, their relative difference was negligible.

RCI derived from the internal consistency estimates was 35 mm for the weighted score and 105 mm for the raw score. Given this dramatic difference, we explored how many patients would be classified as reliably changed. Using the weighted score and the respective RCI, 123 patients would be classified as reliably improved and 20 as reliably deteriorated. Using the raw score and the respective RCI, only 70 patients would be classified as reliably changed and 6 as reliably deteriorated. We concluded that the weighted score, due to a higher precision of measurement, is substantially more sensitive to change.

The clinical cutoff score, determined using Jacobson and Truax's (1991) method "c", was 206 mm for the weighted score and 211 mm for the raw score. In our sample, 80 patients would be classified as clinically improved (i.e., having moved from the clinical range to the non-clinical range) based on the weighted score and the respective cutoff, while 77 would be so using the raw score and the respective cutoff. We deemed this difference negligible.

DISCUSSION

The primary aim of the study was to test the factor structure of the Czech version of the ORS and its invariance between a clinical and non-clinical sample, between two measurement points (pre-therapy vs. post-therapy), and between two forms of administration (paper-and-pencil vs. online).

We supported the hypothesized one-factor structure of the scale. In our study, the best fitting model for both the clinical and the non-clinical samples was a partially

tau-equivalent model, in which the loadings for Item 1 (personal well-being) and Item 4 (overall well-being) were fixed to the same value. To the best of our knowledge, this is the first study to test the factor structure of the ORS; the results therefore cannot be compared to any previous study.

Factor invariance between the clinical and the non-clinical sample was only metric (i.e., item intercepts differed considerably between the samples), which suggests that the scale does not measure exactly the same construct across the respective populations and, thus, it is problematic to directly compare clinical and non-clinical scores. From this point of view, using the "c" criterion to assess clinically significant change (Jacobson & Truax, 1991) seems problematic and should be used with caution. Nevertheless, given the near-to-threshold fit values (SRMR), future studies should further investigate scalar (and strict) invariance between clinical and non-clinical populations and verify this conclusion.

Importantly, the factor structure demonstrated scalar invariance across the course of the treatment (i.e., pretreatment vs. post-treatment). Error variances in items were lower in post-therapy measurement (strict noninvariance), and latent variance increased from pre- to post-therapy, which led to higher reliability of the post-therapy measurement. Referring to the discussion about the meaningfulness of incremental changes in the scores (Sandell & Wilczek, 2016), we can conclude that the ORS can be safely used to assess the therapeutic change in the clinical population because the scale measures the same construct before and after psychotherapy.

We have also found that the factor structure was invariant across two forms of administration (i.e., paper-and-pencil vs. online) in the sense of strict invariance (same item intercepts, loadings, and error variances). The only difference was in the latent means – the participants who filled the questionnaire online had lower mean scores, with a small effect size. We were unable to interpret this effect and suggest investigating it in future studies. We propose treating both forms as equivalent and their scores directly comparable. This finding is relevant, considering that current monitoring systems rely on online administration of the scale.

We discovered that the items considerably differed in the weights they had in producing the latent score. For instance, in the clinical sample, the latent score was almost fully determined by Item 4. Although it might seem that Items 1 to 3 play a negligible role in the clinical population, we do not agree with such an interpretation. They may be necessary in order to "calibrate" a patient before they provide a considered answer to Item 4.

However, the dramatic differences in item weights suggest that it might be more appropriate to compose the total score as a weighted sum of item values. To explore the consequences of this method regarding the psychometric properties of the scale, we conducted a series of secondary analyses. We found that there were no essential differences between the weighted and raw scores in terms of correlations with reference outcome measures and associations with demographic variables. Both versions also yielded similar results in terms of effect sizes in pre-/post-treatment measurement and clinical significance of change. However, the weighted score yielded substantially better estimates of internal consistency, leading to higher measurement precision and a narrower RCI. Consequently, it was much more efficient in detecting a statistically reliable change in patients. Therefore, in situations in which decision-making is based on RCI (such as with the rational method in routine outcome monitoring and feedback systems, Castonguay et al., 2013), the use of the weighted score is preferable. Before implementation into routine practice, however, replication studies are needed to assess the generalizability of the item weights we derived in our study.

We found that the concurrent validity of the Czech ORS with reference instruments, as well as the internal consistency (especially in case of the weighted score), were satisfactory. Our results suggest that the scale is sufficiently sensitive to therapeutic change, which is in line with previous studies (Bieščad, 2007; Miller et al., 2003). Since the scale is expected to be sensitive to week-to-week variations in well-being, the lower values of test-retest reliability can be interpreted as a sign of sensitivity to change.

The generalizability of the results is limited because the clinical sample was composed solely of inpatients and outpatients receiving psychotherapy in daycare centers on a daily basis. It did not include outpatients with less intensive treatment (e.g., once a week). The results also cannot be generalized to patients suffering from psychotic disorders because they were not included either. Therefore, future studies should try to replicate our findings in these populations. Furthermore, the representativeness of the clinical sample may have been biased by a relatively large group of patients who did not consent to be included in the study and the representativeness of the non-clinical and student samples was difficult to assess due to the convenience sampling strategy.

CONCLUSION

We supported the one-factor structure of the ORS in clinical and non-clinical samples. While only metric measurement invariance was found across these two samples, the scale demonstrated scalar invariance across two time points in the clinical sample (pre- and post-therapy) and strict invariance between the paper-and-pencil and online forms of administration in the non-clinical sample. Furthermore, we demonstrated that the measure possesses high internal consistency if the total score is calculated as a weighted sum of items. Test-retest reliability, concurrent validity, and sensitivity to change were satisfactory.

REFERENCES

American Psychological Association (2010). *Ethical principles of psychologists and code of conduct*. http://www.apa.org/ethics/code/principles.pdf

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology, 69*(2), 184-196. doi:10.1037//0022-006x.69.2.184

Behrendt, S. (2014). *Lm.beta: Add standardized regression coefficients to lm-Objects. R package version 1.5-1*. https://CRAN.R-project.org/package=lm.beta

Bieščad, M. (2007). *Aplikácia nástrojov merajúcich výsledky psychoterapie: Porovnanie citlivosti nástrojov merania v jednotlivých oblastiach terapeutickej zmeny* (The application of psychotherapy outcome measures: Comparison of measurement sensitivity in particular areas of therapeutic change). [Unpublished dissertation thesis]. Trnavská univerzita v Trnave.

Biescad, M., & Timulak, L. (2014). Measuring psychotherapy outcomes in routine practice: Examining Slovak versions of three commonly used outcome instruments. *European Journal of Psychotherapy & Counselling, 16*(2), 140-162. https://doi.org/10.1080/13642537.2014.895772

Bringhurst, D. L., Watson, C. W., Miller, S. D., & Duncan, B. L. (2006). The reliability and validity of the Outcome Rating Scale: A replication study of a brief clinical measure. *Journal of Brief Therapy, 5*(1), 23-30.

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

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. (2013). Practice-oriented research: Approaches and applications. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change,* 6th ed (pp. 85-133). John Wiley.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance, structural equation modeling. *A Multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Doran, J. M., Westerman, A. R., Kraus, J., Jock, W., Safran, J. D., & Muran, J. C. (2015, June). *Do all roads lead to Rome? A critical analysis of agreement and divergence in qualitative and quantitative descriptors of change*. Poster presented at the 46th International Annual Meeting of the Society for Psychotherapy Research, Philadelphia, PA, USA.

Hill, C. E., Chui, H., & Baumann, E. (2013). Revisiting and reenvisioning the outcome problem in psychotherapy: An argument to include individualized and qualitative measurement. *Psychotherapy, 50*(1), 68-76. https://doi.org/10.1037/a0030571

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Jabrayilov, R., Emons, W. H. M., de Jong, K., & Sijtsma, K. (2017). Longitudinal measurement invariance of the Dutch Outcome Questionnaire-45 in a clinical sample. *Quality of Life Research, 26*, 1473-1481. https://doi.org/10.1007/s11136-017-1500-1

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.

Janse, P., Boezen-Hilberdink, L., van Dijk, M. K., Verbraak, M. J. P. M., & Hutschemaekers, G. J. M. (2014). Measuring feedback from clients: The psychometric properties of the Dutch Outcome Rating Scale and Session Rating Scale. *European Journal of Psychological Assessment, 30*(2), 86-92. https://doi.org/10.1027/1015-5759/a000172

Juhová, D., Řiháček, T., Cígler, H., Dubovská, E., Saic, M., Černý, M., Dufek, J., & Evans, C. (2018). Česká adaptace dotazníku CORE-OM: vybrané psychometrické charakteristiky [Czech adaptation of the CORE-OM: Selected psychometric properties]. *Československá psychologie, 62*(1), 59-74.

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

Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The treatment outcome package. *Journal of Clinical Psychology, 61*(3), 285-314. https://doi.org/10.1002/jclp.20084

Lambert, M. J. (2015). Progress feedback and the OQ-System: The past and the future. *Psychotherapy, 52*(4), 381-390. https://doi.org/10.1037/pst0000027

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

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57-64. https://doi.org/10.1027/1614-2241.2.2.57

Lyon, A. R., Lewis, C. C., Boyd, M. R., Hendrix, E., & Liu, F. (2016). Capabilities and characteristics of digital measurement feedback systems: Results from a comprehensive review. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(3), 441-466. https://doi.org/10.1007/s10488-016-0719-4

Miller, S. D., Duncan, B. L., Brown, J., Sparks, J., & Claud, D. (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.

O'Reilly, A., Peiper, N., O'Keeffe, L., Illback, R., & Clayton, R. (2016). Performance of the CORE-10 and YP-CORE measures in a sample of youth engaging with a community mental health service. *International Journal of Methods in Psychiatric Research, 25*(4), 324-332. https://doi.org/10.1002/mpr.1500

Prescott, D. S., Maeschalck, C. L., & Miller, S. D. (Eds.) (2017). *Feedback-informed treatment in clinical practice: Reaching for excellence.* American Psychological Association.

Rice, K. G., Suh, H., & Ege, E. (2014). Further evaluation of the Outcome Questionnaire-45.2. *Measurement and Evaluation in Counseling and Development, 47*(2), 102-117. https://doi.org/10.1177/0748175614522268

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

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

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514. https://doi.org/10.1007/BF02296192

Schauenburg, H., & Strack, M. (1999). Measuring psychotherapeutic change with the Symptom Checklist SCL 90 R. *Psychotherapy and Psychosomatics, 68*, 199-206. https://doi.org/10.1159/000012333

Scott, K., & Lewis, C. C. (2015). Using measurement-based care to enhance any treatment. *Cognitive and Behavioral Practice, 22*(1), 49-59. https://doi.org/10.1016/j.cbpra.2014.01.010

Zatloukal, L., Žákovský, D., Věžník, M., Řiháček, T., & Tkadlčíková, L. (2006). *Česká verze škál ORS a SRS [The Czech version of the ORS and the SRS]*. http://www.dalet.cz/Clanky/scales-CZ.pdf

SOUHRN

Česká verze Outcome Reting Scale: vybrané psychometrické charakteristiky

Cíle. Outcome Rating Scale (ORS) je velmi krátká sebehodnotící škála určená k měření změn během psychoterapie. Cílem této studie bylo ověřit (a) faktorovou strukturu ORS, (b) invarianci měření mezi klinickým a neklinickým vzorkem, mezi hodnocením před léčbou a po léčbě (v rámci klinického vzorku) a mezi online a papírovou formou škály (v rámci neklinického vzorku), (c) souběžnou validitu s dalšími nástroji na měření výsledku psychoterapie a (d) citlivost na terapeutickou změnu.

Vzorek. Studie se zúčastnilo $N = 256$ pacientů, $N = 210$ neklinických respondentů a $N = 89$ studentů. Pacienti vyplnili ORS před psychoterapií a po ní.

Statistická analýza. Faktorová struktura a invariance měření byly ověřovány pomocí konfirmační faktorové analýzy. Souběžná validita a stabilita v čase byly posuzovány pomocí korelační analýzy. Citlivost na změnu byla hodnocena pomocí indexu spolehlivé změny a velikosti účinku.

Výsledky. Byla potvrzena jednodimenzionální struktura škály. Nejvhodnějším modelem byl částečně tau-ekvivalentní model s náboji první a čtvrté položky fixovanými na stejnou hodnotu. Zatímco mezi klinickým a neklinickým vzorkem byla potvrzena pouze metrická invariance, mezi hodnocením před a po léčbě škála vykazovala skalární invarianci a mezi online a papírovou formou přísnou invarianci. Vnitřní konzistence i souběžná validita byly uspokojivé. Citlivost na změnu psychoterapie byla adekvátní. Vnitřní konzistence a citlivost na změnu se zvýšila, pokud byl celkový skór počítán jako vážená suma položek.

Omezení studie. Vzorky nebyly reprezentativní.