European Journal of Psychotherapy & Counselling
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rejp20

Measuring psychotherapy outcomes in routine practice: Examining Slovak versions of three commonly used outcome instruments

Matus Biescad[a] & Ladislav Timulak[b]

[a] Department of Psychology, Trnava University, Trnava, Slovakia

[b] School of Psychology, Trinity College Dublin, Dublin, Ireland

Published online: 19 Mar 2014.

To cite this article: Matus Biescad & Ladislav Timulak (2014): Measuring psychotherapy outcomes in routine practice: Examining Slovak versions of three commonly used outcome instruments, European Journal of Psychotherapy & Counselling, DOI: 10.1080/13642537.2014.895772

To link to this article: http://dx.doi.org/10.1080/13642537.2014.895772

European Journal of Psychotherapy & Counselling, 2014
http://dx.doi.org/10.1080/13642537.2014.895772

Routledge
Taylor & Francis Group

Measuring psychotherapy outcomes in routine practice: Examining Slovak versions of three commonly used outcome instruments

Matus Biescad[a]* and Ladislav Timulak[b]

[a]*Department of Psychology, Trnava University, Trnava, Slovakia;* [b]*School of Psychology, Trinity College Dublin, Dublin, Ireland*

(*Received 28 October 2013; accepted 25 January 2014*)

This study examined the Slovak mutations of three outcome measures for routine practice i.e. the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM), the Outcome Questionnaire – 45 (OQ-45), the Outcome Rating Scale (ORS), and one control measure the Symptom Checklist 10 Revised (SCL-10R), with regard to their concordance or differences in outcome classification of pre-post change, when used by the same patients and when the criteria used for establishing recovery and improvement status are based on the same sample. *Method*: Non-clinical (252) and clinical (202) samples were used for the standardisation of all instruments. A portion of the clinical participants ($N = 140$) completed all measures at the end of their treatment. *Results*: The CORE-OM, and the SCL-10R indicated a higher number of recovered and improved clients. With regard to the pre-post differences as expressed in the effect size, the CORE-OM showed the highest pre-post difference (pre-post effect size .98), followed by the ORS (.87), the SCL-10R (.83) and finally with the OQ-45 (.69). *Conclusion*: Even very similar instruments developed on the basis of similar theoretical conceptualisations and empirical findings may report different pre-post outcomes.

Keywords: measuring outcomes in routine practice; concordance in outcome classification of pre-post-treatment change; Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM); Outcome Questionnaire – 45 (OQ-45); Outcome Rating Scale (ORS)

Este artículo examina las transformaciones eslovacas en tres instrumentos de medida en la práctica diaria de la psicoterapia, es decir, los resultados clínicos en la práctica diaria: Medida de Resultados; el Cuestionario de Resultados-45; la Escala de Evaluación de Resultados y una medida de control, el SCL-10R en relación con sus concordancias o diferencias en la clasificación de resultados de pre-post tratamiento cuando son usados por el mismo paciente, y cuando el criterio utilizado para establecer la recuperación y el mejoramiento de la situación se basa en la misma muestra. Método: Se utilizaron muestras clínicas y no-clínicas para la estandardizac-

*Corresponding author. Email: mbiescad@gmail.com

ión de todos los instrumentos. Una parte de los participantes clínicos (N=140) completó todas las medidas al final de su tratamiento. Resultados: el CORE-OM y el SCL-10R indicaron una cifra alta de clientes que habían mejorado y recobrado su estado normal. En relación con el efecto del tamaño de la muestra, el CORE-OM mostró la más alta pre-post diferencia (pre-post efecto .98) seguido por el ORS (.87), el SCL-10R (.83) y finalmente el OQ (.69). Conclusión: Incluso instrumentos similares pueden indicar diferentes pre-post resultados.

Palabras clave: medida de resultados en la práctica diaria; concordancia en la clasificación de los resultados de cambio pre y post tratamiento; resultados clínicos de las medidas delos resultados en la evaluación diaria (CORE-OM); cuestionario de resultados (OQ-45); escala de evaluación de resultados

Cette étude examine les versions slovaques de trois instruments de mesure utilisés quotidiennement dans la pratique de la psychothérapie, à savoir le CORE-OM [Clinical Outcomes in Routine Evaluation – Outcome Measure], le OQ-45 [the Outcome Questionnaire – 45], and le ORS [Outcome Rating Scale] ainsi qu'une mesure de contrôle, le SCL-10R quant à leur concordance ou divergence dans la classification des résultats pré et post changement, lorsqu'ils sont utilisés par les mêmes patients et lorsque les critères utilisés pour établir les conditions de la guérison et de l'amélioration sont basés sur le même échantillon. Méthode: des échantillons non-cliniques (252) et cliniques (202) ont été utilisés pour la standardisation des tous les instruments. Une partie des participants cliniques (N=140) ont rempli toutes les mesures à la fin de leur traitement. Résultats: le CORE-OM et le SCL-10R indiquent la guérison et l'amélioration d'un plus grand nombre de clients. Concernant les différences pré-post comme exprimées dans l'ampleur de l'effet, le CORE-OM montre la plus grande différence pré-post (ampleur d'effet pré-post .98), suivi par le ORS (.87) et le SCL-10R (.83) et enfin le OQ-45 (.69). Conclusion: Des instruments, pourtant très similaires, développés selon des conceptualisations théoriques similaires et des résultats empiriques, peuvent donner des résultats pré-post différents.

Mots-clés: mesure des résultats dans la pratique quotidienne; concordance dans la classification des résultats pré- post traitement; CORE-OM; OQ-45; ORS;

The measurement of outcomes in routine practice is becoming more and more common (Barkham & Mellor-Clark, 2003). There are several reasons for this which include: the ethical responsibility of monitoring the course of treatment; the possibility of altering and improving treatment on the basis of on-going outcomes; an accountability to all stakeholders (e.g. clients or insurance companies), and contributions to research and general knowledge (in case of large naturalistic samples), etc. (cf. Cone, 2001; Lambert, 2010; Lambert & Finch, 1999; Ogles, Lambert, & Fields, 2002). Some of those reasons are controversial (Evans, 2012a).

Instruments being developed are suitable for use in both sophisticated research studies such as randomised clinical trials and in everyday practice (see Ogles et al., 2002). Therefore, these measures must have sound psychometric properties, be clinically relevant and be adequately sensitive in order to capture any change that occurs in therapy. As a result, they typically focus on the areas that research shows are affected by psychotherapy, such as general sense of well-being, psychopathological symptoms and interpersonal functioning (Howard, Lueger, Maling, & Martinovich, 1993; Howard, Moras, Brill, Martinovich, & Lutz, 1996). It is necessary that these measures are also clinically practical, that is that they are user friendly for both the client and the therapist, they are sensitive to change and they focus on clinically relevant issues (such as suicide risk). Furthermore, they should be universal in order to be used with a variety of clients, can be used repeatedly, easily scored and transtheoretical (i.e. not used only by one school of psychotherapy) (cf. Barkham et al., 1998; Evans et al., 2000; Lambert et al., 2004; Lambert, Okiishi, Finch, & Johnson, 1998).

Several measures have recently been developed in order to assess therapy outcomes in routine practice (see Ogles et al., 2002). They are usually accompanied by web-based or computer-based programmes which provide an immediate score. These scores can then be benchmarked against large datasets which provide a comparison between the evaluated clients and the relevant referential group. These measure are often used for tracking the session-by-session outcomes of the client, against the curves of session-to-session improvements of successfully recovered or improved clients (Lambert, Hansen, & Finch, 2001; Lambert et al., 2003, 2001; Miller, Duncan, Brown, Sparks, & Claud, 2003; Miller, Duncan, Sorrell, & Brown, 2005).

In this study, we wanted to examine Slovak mutations of three commonly used measures assessing the outcomes in routine practice. Routine evaluation of treatment outcomes is expected to be a requirement for the majority of service providers, thus we wanted to contribute to the debate about the ways in which the evaluation of outcomes in routine practice may be implemented. We selected three instruments that were specifically developed to assess the outcomes of psychotherapy. These instruments are widely used for the evaluation of outcomes in routine practice in other jurisdictions. All have online versions with extensive benchmark data that allows for an on-going monitoring of outcomes against the course of successful treatment of clients with similar initial characteristics.

Specifically, we wanted to examine the Slovak versions of the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM; Barkham et al., 2001), the Outcome Questionnaire – 45 (OQ-45; Lambert et al., 2004) and the Outcome Rating Scale (ORS; Miller et al., 2003). Principally, we wanted to test their concordance or differences in outcome classification of pre-treatment vs. post-treatment change, when used by the same clients and when the criteria used for establishing recovery and improvement status are based on the same referential samples.

The CORE-OM (CORE System Group, 1998) is a self-report instrument that was developed to be used with a variety of clients, undergoing a variety of treatments, in a variety of settings (Barkham et al., 1998, 2001; Barkham, Mellor-Clark, Connell, & Cahill, 2006; Evans, 2012b; Evans et al., 2000). The instrument consists of 34 items, covering four domains (well-being, problems [psychological symptoms], functioning [interpersonal and social functioning] and risk behaviour [risk to self and to others]). The instrument has good psychometric properties (Evans et al., 2002) and correlates highly with other established measures such Beck Depression Inventory or Hamilton Rating Scale of Depression (Cahill et al., 2006). It has proved to be sensitive to therapeutic change in different types of settings (Evans et al., 2000). It can be also found in an abbreviated form consisting of 18 (Barkham et al., 2001), respectively ten and five items (Connell & Barkham, 2007) and has been translated into several languages (e.g. Elfström et al., 2012; Palmieri et al., 2009; Skre et al., 2013). The psychometric qualities along with initial normative and clinical population data have previously been established in the Slovak version, (Gampe, Bieščad, Balúnová-Labaničová, Timuľák, & Evans, 2007).

The OQ-45 (Lambert et al., 2004) is an instrument developed to monitor the outcome of therapy over time, across the relevant psychopathological symptoms and the social roles that the client engages in. It can be used in managed care in a variety of agencies (Lambert et al., 1996; Lambert & Finch, 1999). The OQ-45 is a 45-item self-report measure that focuses on three areas of therapeutic change: symptom distress, interpersonal relations and social role performance as conceptualised by Lambert (1983). The instrument has adequate psychometric properties (Lambert et al., 1996) and correlates highly with other theoretically similar measures such as Zung Depression Scale, State-Trait Anxiety Inventory and Inventory of Interpersonal Problems, BASIS-32 (Doerfler, Addis, & Moran, 2002; Lambert et al., 1996). The instrument was also satisfactorily examined to measure its sensitivity to therapeutic change (Lambert et al., 2004; Vermeersch, Lambert, & Burlingame, 2000; Vermeersch et al., 2004). The OQ-45 also has shortened forms, 10-item OQ-45 and 30-item Life Status Questionnaire (LSQ-30). The OQ-45 was translated to several languages, among them German (Lambert, Hannöver, Nisslmüller, Richard, & Kordy, 2002), Portuguese (Machado & Klein, 2006), Italian (Lo Coco et al., 2006) and Dutch (de Jong et al., 2007). The Slovak version was established as a part of the study presented here and previously presented by Biescad and Timulak (2006).

The final measure in this study, The ORS (Miller et al., 2003; Miller, Duncan, Sorrell, et al., 2005) was developed specifically for measuring and monitoring the progress of therapy and outcome of therapy. The main goal in developing the ORS was to strike a balance between the reliability and validity of the measure with the simplicity in its administration and scoring (Miller, Duncan, & Hubble, 2004, 2005). The ORS (International Centre for Clinical Excellence, 2013) is an ultra-brief visual-analogue scale, conceptually based on OQ-45, measuring three areas (personal well-being, close interpersonal relationships, and social functioning at work, school or with friends) along with a measure of well-being. The four areas are represented by four horizontal visual scales 10 cm long on which the client expresses how he or she is currently functioning (closer to the right side means better). The measure has adequate psychometric properties and while it correlated moderately with the OQ-45 (Miller et al., 2003) its test–retest reliability was lower. The measure was translated into several languages; however, information about the psychometric qualities of those translations has not yet been reported. The Slovak version was prepared by the first author of this paper and the psychometric qualities were reported in Zvarová and Bieščad (2007).

The aim of the current study was to examine the potential usefulness of these three instruments for the Slovak mental health services. Specifically, we wanted to explore their practicality for routine use, particularly from the perspective of their concordance or differences in outcome classification of therapeutic change. All three measures examined here were selected as they were developed to be particularly sensitive to change as a result of psychotherapy. This sensitivity to therapeutic change is an important concept as many measures that were traditionally used for psychological assessment (such as MMPI or Rorschach) are not particularly sensitive to the changes achieved by psychotherapy (Lambert & Hill, 1994).

Each of the original instruments (i.e. English language versions) were standardised using different normative and clinical samples. This makes it difficult to fully compare their concordance or differences in outcome classification of therapeutic change, without re-standardisation on the same clinical and non-clinical samples. This current study is unique as reliable change indexes (RCIs) and cut-off scores will be established for each instrument using the same normative and clinical samples. This will enable a direct comparison between these instruments, which were specifically developed to measure therapy outcomes, in terms of their concordance or differences in outcome classification of therapeutic change among the same sample of clients.

Accordingly, this study measured these three instruments in terms of their concordance or differences in outcome classification of therapeutic change. Specifically, we examined the pre-post outcome, as expressed in the effect size, using the same sample of clients undergoing mental health treatment, for each instrument. We also wanted to examine whether, a) each of the measures indicated the same percentage of clients as improved (reliably changed) and recovered (reliably changed and in non-clinical range) and b) if each instrument indicated the same clients as improved and recovered.

Method

Participants

Two samples, non-clinical ($n = 252$) and clinical ($n = 202$), were used for the standardisation of the Slovak mutations of all instruments used in the study. Participants were each required to complete four measures (apart from the CORE-OM, OQ-45, and ORS, we used the 10-item version of the Symptom Checklist – 90 (SCL-90; Derogatis, Lipman, & Covi, 1973), the Symptom Checklist 10 Revised (SCL-10R; Rosen et al., 2000) as a control measure and which will be presented below). There was some attrition and so we report only on those participants who completed all four measures.

A combination of convenience and snowball sampling was used to recruit non-clinical participants. This sample, who were voluntary participants were mainly students, attending evening classes at a university with which the first author was affiliated. A single condition was applicable to this sample, namely

Table 1. Demographic characteristics of the clinical and non-clinical sample of participants that filled in all four measures for referential purposes.

Characteristics		Non-clinical sample ($n = 237$)	Clinical sample ($n = 173$)
Gender	Male	111 (46.8%)	79 (45.7%)
	Female	126 (53.2%)	94 (54.3%)
Age (in years)	Mean	30.70	40.49
	Standard deviation	13.08	12.44
	Range	18–79	18–80
Marital status	Single	161 (67.9%)	63 (36.4%)
	Married	62 (26.2%)	75 (43.4%)
	Divorced	4 (1.7%)	22 (12.7%)
	Widow/er	3 (1.3%)	7 (4.0%)
	Cohabitating	5 (2.1%)	5 (2.9)

that they had not been in psychological or psychiatric care within the last 12 months.

Table 1 presents the demographic details of the non-clinical participants (*N* = 237) who completed all measures being studied. The clinical participants who completed all four measures (*N* = 173) were recruited from two inpatient psychiatric clinics (*N* = 135) and from 6 outpatient facilities (*N* = 38). The demographic details of the clinical participants are presented in Table 1. The official primary diagnoses (using International Classification of Diseases version 10) as assessed by the treating clinician of those participants were: depression (*N* = 51; 28.90%), alcohol abuse (*N* = 51; 28.90%), anxiety disorder (*N* = 25; 14.45%), schizophrenia (*N* = 13; 7.51%) and others, such as personality disorders, eating disorders, other addictions, etc. (~15%). The data from all non-clinical and clinical participants who completed the four measures were used for establishing cut-off scores, distinguishing between the clinical and non-clinical populations. A portion of clinical participants (*N* = 140) also completed the measures when discharged at the end of their treatment, and these data were used for pre-post comparison between the measures (for demographic data see Table 2). The majority of these participants (87%) underwent a routine inpatient treatment consisting of a combination of psychological therapy (in 93% of cases

Table 2. Clinical sample used for pre-post measurement – demographic data.

Characteristics		$n = 140$
Gender	Male	66 (47.1%)
	Female	74 (52.9%)
Age (in years)	Mean	41.6
	Standard deviation	12.7
	Range	19–79
Marital status	Single	48 (34.3%)
	Married	62 (44.3%)
	Divorced	20 (14.3%)
	Widow/er	6 (4.3%)
	Cohabitating	4 (2.9%)

– predominantly eclectic and cognitive-behavioural therapy) and pharmacotherapy (91%). These treatments were provided by teams of practitioners (mainly psychiatrists and psychologists). Additionally, treatment could also include other psychosocial interventions such as occupational therapy.

A further sample consisting of MSc students ($N = 49$, 45 female and 4 male; age range 19–30; mean 22 years [SD = 2.4 years]) was used for assessing test–retest reliability for all four measures.

Measures

CORE-OM (CORE System Group, 1998) – the Slovak version

This is a 34-item self-report measure examining four domains of distress (well-being – W, problems – P, functioning – F, and risk – R) amenable to change by psychotherapy. Each item is scored on a scale of 0–4 and the overall score is stated as the average score per item. The higher the client's score, the greater their distress.

The Slovak language instrument used in this study was the third version and was prepared using the second translation (Gampe et al., 2007) which had been prepared by two translators. This version was compared with three independent translations (translators were psychologists, except for one translator who was a linguist and thus professionally a lay person). The final wording of the Slovak CORE-OM was then established by the meeting of all translators (including those involved in the original translations) with one of the authors of the original CORE-OM (Chris Evans). The data on reliability of the second version of the translation were also taken into account when concluding the final wording of the items (for more on the process of translation see Biešcad (2007)).

The measure was correlated with other measures of psychopathology. It was positively correlated with the overall score on the Beck Depression Inventory (Beck, Epstein, Brown, & Steer, 1988; .84 for the non-clinical and .87 for the clinical sample), and with the Global Severity Index of the SCL-90 (Derogatis et al., 1973; .78 for the non-clinical and .83 for the clinical sample). It was negatively correlated with the Rosenberg Self-Esteem Scale (Rosenberg, 1965; −.58 for the non-clinical and −.74 for the clinical sample) (all questionnaires were Slovak versions; Biešcad, 2007). The internal consistency, Cronbach's α, for the current sample (clinical and non-clinical samples combined) was .96. The two-week test–retest reliability obtained on the sample of students ($N = 49$) and used for the calculation of the RCI used in the current study (see below) was .724. Consequently, the RCI was established with the value of .70 (for the calculation see the procedure section below). The cut-off score distinguishing between the clinical and non-clinical population was set at 1.04 for men and at 1.29 for women (for the calculation see the procedure section below).

The OQ-45 (Lambert et al., 2004) – the Slovak version

This is a 45-item self-report measure examining three domains of distress (symptom distress, interpersonal relations, and social role performance) amenable to change by psychotherapy. Each item is scored on a scale of 0–4 and the main final score is the sum of all scores. The higher the client's score, the greater their distress. The instrument was translated into Slovak language. The translation was initially prepared by four translators, who first worked independently and then consensually prepared the agreed version. This version was then administered to 78 Psychology students and along with an internal consistency reliability, the data on comprehensibility of individual items were gathered (participants could comment on any item of which the meaning was not clear to them). Two more translators (one of which was a non-psychologist) were then added to the team of translators and the final version of the items was agreed (for more on the process of translation see Biešcad (2007)).

The measure was correlated with other measures of psychopathology. It correlated positively with the overall score on the Beck Depression Inventory (.75 for the non-clinical and .87 for the clinical sample) and with the Global Severity Index of the SCL-90 (.77 for the non-clinical and .92 for the clinical sample). It correlated negatively with the Rosenberg Self-Esteem Scale (−.58 for the non-clinical sample) (all questionnaires were the Slovak versions; Biešcad, 2007). The internal consistency, Cronbach's α, for the current sample (clinical and non-clinical samples combined) was .95. The two-week test–retest reliability obtained on the sample of students ($N = 49$) was used for the calculation of the RCI in the current study (see below) was .82. Consequently, the RCI was established with the value of 21.02 (for the calculation see the procedure section below). The cut-off score distinguishing between the clinical and non-clinical population was set at 52.58 for men and at 62.13 for women (for calculation see the procedure section below).

The ORS (Miller et al., 2003; Miller, Duncan, Sorrell, et al., 2005) – the Slovak version

The ORS is an ultra-brief visual-analogue scale measuring three areas (personal well-being, close interpersonal relationships, and social functioning at work, school or with friends) and a general sense of well-being. The four areas are represented by four visual scales, 10 cm long, on which the client expresses how he or she is functioning (closer to the right means better). The overall score is obtained by the sum of all four scales (score expressed in millimetres, each scale is 0–100 mm and the overall score up to 400 mm). A higher the score indicates higher functioning. The translation was initially prepared by four translators, who first worked independently and then consensually prepared the agreed version. This version was then administered to a variety of participants ($N = 93$) and apart from internal consistency, reliability data on comprehensibility was gathered (participants could comments on any items the meaning of which was not fully clear). Two more translators (one of which was a non-psychologist) were then added to the team of translators and the final version of the items was agreed (for more on the process of translation see Biešcad (2007)).

The measure was correlated with other measures of psychopathology. It correlated with the overall score on the Beck Depression Inventory (−.73 for the non-clinical and −.74 for the clinical sample), with the Global Severity Index of the SCL-90 (−.59 for the non-clinical and −.67 for the clinical sample) and negatively with the Rosenberg Self-Esteem Scale (.38 for the non-clinical sample) (all questionnaires were Slovak versions; Biešcad, 2007). The internal consistency, Cronbach's α, for the current sample (clinical and non-clinical samples combined) was .87. The two-week test–retest reliability obtained on the sample of students ($N = 49$) that was used for the calculation of the RCI in the current study (see below) was .735. Consequently, the RCI was established with the value of 111.6 (for the calculation see the procedure section below). The cut-off score distinguishing between the clinical and non-clinical population was set at 251.3 for men and at 228.1 for women (for calculation see the procedure section below).

The SCL-10R (Rosen et al., 2000) – the Slovak version

The SCL-10R is a measure which uses ten items of the SCL-90 (Derogatis et al., 1973) and is one of the brief versions of the SCL-90 (e.g. SCL-25, Holi, 2003; SCL-9, Klaghofer & Brähler, 2001; SCL-10, Nguyen, Attkisson, & Stegner, 1983; SCL-6, Rosen et al., 2000). We chose this measure as another benchmark for the primarily studied measures as it can also be used for monitoring and assessing outcome in routine practice. It is based on a well-known and widely used measure the SCL-90. The SCL-90 is a 90-item self-report instrument measuring current psychological distress. It covers items across nine dimensions (somatisation, obsessions-compulsions, interpersonal sensitivity, depression, anxiety, hostility, paranoid ideation, phobias, and psychoticism) and is used for the psychopathological assessment, but often also as a measure of psychotherapy outcome.

Rosen et al. (2000) prepared the SCL-10R as a 10-item instrument on the basis of 12 factor analyses of the SCL-90. The items included in the SCL-10 loaded highly on the primary factor of the overall SCL-90 (Global Severity Index) and represented as many as possible of the nine dimensions of the SCL-90. Each item is scored on a scale of 0–4 and as the overall score is stated as the average score per item. The higher the average score, the greater the client's distress. When tested against the Slovak version of the SCL-90, the correlation was .95 and internal consistency .90 (Biešcad, 2004). The Slovak version of SCL-10 was prepared by 4 translators, who first worked independently and then consensually agreed the final version. When concluding the final version, the translators also compared their wording with the wording in the Czech version of the SCL-90 (Boleloucký, 1993) as the two languages are similar.

The SCL-10 was correlated with other measures of psychopathology and correlated positively with the overall score on the Beck Depression Inventory

(.70 for the non-clinical sample), with the Global Severity Index of the SCL-90 (.80 for the non-clinical sample) and with the Rosenberg Self-Esteem Scale (−.53 for the non-clinical sample) (all questionnaire were Slovak; Bieščad, 2007). The internal consistency, Cronbach's α, for the current sample (clinical and non-clinical samples combined) was .98. The two-week test–retest reliability obtained on the sample of students ($N = 49$) that was used for the calculation of the RCI in the current study (see below) was .801. Consequently, the RCI was established with the value of .64 (for the calculation see the procedure section below). The cut-off score distinguishing between the clinical and non-clinical population was set at .85 for men and at 1.26 for women (for calculation see the procedure section below).

Procedure

After the measures were translated (Bieščad, 2007), all four (CORE-OM, OQ-45, ORS, and SCL-10R) were administered to the same non-clinical sample ($n = 252$) and the same clinical sample ($n = 202$). To provide a counterbalance to any order effects, the participants completed the measures in a random order (we used a balanced Latin square randomisation). Participants were informed that their participation was voluntary and that there would be no negative consequences should they wish to withdraw from the study at any point. Participants completed an Informed Consent form at the beginning of the research. Details of the recruitment were presented in the Participants section. A small number of participants did not complete all four measures and casewise deletion was applied. The final non-clinical sample consisted of two hundred and thirty-seven ($N = 237$) participants and one hundred and seventy-three ($N = 173$) participants were included in the clinical sample (see Table 3 for the numbers of missing data). This may mean that some of the instruments were possibly more difficult to fill in (this was the case for the OQ-45 in the clinical sample and for the ORS in the clinical and non-clinical sample).

The data obtained from the participants, who completed all four instruments, were used to calculate the clinical cut-off scores, differentiating between the clinical and non-clinical population for each instrument (see the section Measures where the actual cut-off scores are stated). Since the distributions of scores of the clinical and non-clinical samples were overlapping for each instrument and since we had both normal and clinical samples at our disposal, Jacobson and Truax's (1991) formula *c* for calculating the cut-off score was

Table 3. The number of missing data for each of the instruments.

	Non-clinical sample ($n = 252$)	Clinical sample ($n = 203$)
CORE-OM	2	3
OQ-45	2	17
ORS	11	18
SCL-10R	2	1

used. The formula for calculating c stands as follows (Jacobson & Truax, p. 13):

$$c = \frac{\mathrm{SD}_0 M_1 + \mathrm{SD}_1 M_0}{\mathrm{SD}_0 + \mathrm{SD}_1}$$

where SD_0 is the standard deviation and M_0 the mean of the normal sample (population) and SD_1 is the standard deviation and M_1 the mean of the clinical sample (population).

A sample of students ($N = 49$) who completed all four instruments (for more see the section Participants) was used for calculating test–retest reliability (the actual numbers are stated in the section Measures). The test–retest reliability was then used to calculate the RCI (the minimum amount of change that could not be attributed to the error of measurement). We followed Jacobson and Truax's (1991) recommendation for the calculation and used a variation of the original Jacobson and Truax's formula that was recommended by Wiger and Solberg (2001, p. 148):

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

where 1.96 represents the Z score for the 95% confidence interval, SD is the standard deviation of the normal sample (population) distribution and rel is test–retest reliability.

Thus we obtained the cut-off scores and RCIs for all four measures from the same sample of participants who represented both non-clinical and clinical populations. These were then used (see the next paragraph) to compare the number of participants recovered, (pre-post difference reliably changed and the participant starting treatment in the clinical and finishing in the non-clinical population, as determined by the cut-off score), improved (pre-post difference reliably changed but the participant still in the clinical population), not changed (pre-post difference lower than RCI) or deteriorated (pre-post difference reliably worsened) on each of the instruments.

A subsample of the clinical sample ($N = 140$) completed all four instruments at the beginning of treatment and at their discharge. As mentioned in the Participants section, this subsample was mainly inpatient, with some patients from outpatient facilities. The treatment length varied from 4 days to 41 weeks. All participants underwent some form of psychotherapy (individual [eclectic, cognitive-behavioural, psychodynamic, humanistic], group therapy or a combination of treatments), although the inpatient patients typically also underwent some other form of treatment (mainly pharmacotherapy). The pre-post data from this sample were used to compare the concordance or differences in outcome classification of therapeutic change between the four measures. The pre-post change as expressed in an effect size (Cohen's d) was determined for each of the instruments, as well as the number of recovered, improved, not changed, and deteriorated patients as shown on each of the measures. The match between the end of treatment status of the client on each of the instruments (recovered, improved, not changed, deteriorated), was also established, in order to determine whether the instruments judged a specific client's status similarly.

Data analysis

We conducted three main analyses. We compared the effect sizes of each of the instruments using Cohen's d (Cohen, 1988) using the following formula:

$$d = \frac{\bar{x}_{\text{pre}} - \bar{x}_{\text{post}}}{\sqrt{\frac{S_{\text{pre}}^2 - S_{\text{post}}^2}{2}}}$$

where $\bar{x}_{\text{pre}}$ and $\bar{x}_{\text{post}}$ represent the arithmetic mean at the beginning and at the end of treatment and S_{pre} and S_{post} represent the standard deviation at the beginning and at the end of treatment.

We also examined the association of the instrument used and the number of recovered, improved, not changed, and deteriorated clients using χ^2. We wanted to determine whether any of the instruments identified more clients as recovered and improved when compared to the other instruments. Finally, we also wanted to examine the level of match between the four instruments in identifying the clients as recovered, improved, not changed, and deteriorated. For that we used coefficient κ (although we also examined simple indexes of agreement p_0).

Results

First, we wanted to look at the pre-post change on clients who completed the fours instruments pre and post treatment. The pre-post change, as measured by a paired T-test, on all measures and all their subscales reached statistical significance (see Table 4). We were, however, primarily interested in the effect sizes. As can be seen in Table 4, the CORE-OM's overall score's effect size was .98, while the ORS's was .87, the SCL-10R's .83, and the OQ45's was .69. This would suggest for instance, that the difference between the CORE-OM and OQ-45 pre-post change, measured on the same clients, would be almost one-third of the standard deviation in the direction of a bigger pre-post change as measured by the CORE-OM (perhaps suggesting a greater sensitivity to change for that measure and for this sample). The ORS and the SCL-10R achieved a pre-post effect size somewhere between the CORE-OM and the OQ-45.

Another aspect of measuring the concordance or differences in outcome classification of therapeutic change is to establish how many clients, according to a given measure at the end of treatment, recovered, improved, did not change, and/or deteriorated. To answer this question we used RCIs and the cut-off scores that were established earlier (see the Method section) using the referential clinical and non-clinical data as well as test–retest reliability. First, as we were aware that not all clients were necessarily in the clinical range of the measure at the beginning of the treatment, we examined how many clients according to each of the measure met the criteria for clinical caseness. Table 5

Table 4. Comparison of the average pre-post score for all fours measures on the same patients ($N = 140$).

	Pre-treatment score *M* (SD)	Post-treatment score *M* (SD)	Cohen's *d*	*t*(df)
CORE-OM				
W	2.40 (1.03)	1.56 (.89)	.87	9.91 (139)***
P	2.19 (.97)	1.35 (.84)	.93	10.76 (139)***
F	1.80 (.75)	1.18 (.70)	.85	9.58 (139)***
R	.63 (.75)	.23 (.41)	.66	7.24 (139)***
Overall score-R	2.05 (.81)	1.31 (.74)	.95	10.81 (139)***
Overall score	1.80 (.74)	1.12 (.65)	.98	11.13 (139)***
OQ-45				
SD	46.72 (18.57)	34.04 (18.27)	.69	10.14 (139)***
IR	17.23 (7.97)	13.61 (7.36)	.47	6.99 (139)***
SR	14.69 (5.91)	11.44 (5.19)	.58	7.21 (139)***
Overall score	78.64 (28.53)	59.09 (27.75)	.69	9.95 (139)***
ORS				
Individually	40.90 (27.23)	67.15 (24.66)	1.01	−11.52 (139)***
Interpersonally	55.55 (29.79)	67.16 (27.55)	.40	−5.67 (139)***
Socially	49.79 (28.17)	66.89 (24.73)	.65	−8.87 (139)***
Overall	42.82 (28.17)	66.92 (25.19)	.90	−10.47 (139)***
Overall score	189.06 (91.31)	268.11 (90.04)	.87	−11.17 (139)***
SCL-10R				
Overall score	1.73 (.92)	1.03 (.75)	.83	10.74 (139)***

$^{***}p < .001$.

Notes: CORE-OM – Clinical Outcome in Routine Evaluation – Outcome Measure; W – well-being; P – problems/symptoms; F – functioning; R – risk; Overall score–R – overall score without risk items; OQ-45 – Outcome Questionnaire – 45; SD – symptom distress; IR – interpersonal relationships; SR – social role; ORS – Outcome Rating Scale; SCL-10R – Symptom Checklist – 10R.

shows that between 70 and 78% of the clients according to the various measures met the criteria for clinical caseness. Table 5 also shows how many of the participants met the criteria at the end of the treatment. As can be seen, the number of clients that moved to the non-clinical range for the CORE-OM was 30.7% (calculated as the difference between the number of clinically distressed patients at the beginning of the treatment minus the number of the clinically distressed patients at the end of the treatment), for OQ-45 it was 23.6%, for the ORS it was 32.8% and for the SCL-10R it was 34.2%. The χ^2 test

Table 5. Categorisation of participants to clinical vs. non-clinical populations ($N = 140$).

	Pre-treatment		Post-treatment	
	Non-clinical	Clinical	Non-clinical	Clinical
CORE-OM	33 (23.6%)	107 (76.4%)	76 (54.3%)	64 (45.7%)
OQ-45	31 (22.1%)	109 (77.9)	64 (45.7%)	76 (54.3%)
ORS	41 (29.3%)	99 (70.7%)	87 (62.1%)	53 (37.9%)
SCL-10R	39 (27.9%)	101 (72.1%)	87 (62.1%)	53 (37.9%)

Notes: CORE-OM – Clinical Outcome in Routine Evaluation – Outcome Measure; OQ-45 – Outcome Questionnaire – 45; ORS – Outcome Rating Scale; SCL-10R – Symptom Checklist – 10R.

examining the association between the measure and the number of participants that moved from the clinical to non-clinical populations at the end of the treatment suggested ($\chi^2 = 13.01$; $p = .162$) that there is no significant association between the measures and the movement from non-clinical to clinical range.

When we examined the differences between the categories of recovered, improved, not changed, and deteriorated clients at the end of the treatment, using the RCIs and cut-off scores, we established, we could see that the CORE-OM and the SCL-10R showed more clients as recovered and/or improved (around 50%) than the OQ-45 (around 40%) and the ORS (around 33%). Indeed, the χ^2 examination of the association between the measure used and the status of the client at the end of the treatment suggested ($\chi^2 = 19.43$; $p = .021$) that there is a significant association between the measures (see also Table 6).

When we looked only at the clients that started treatment in the clinical range, we found that while for the CORE-OM and the SCL-10R it was around 36–37% of the clients that did not change or got worse, for the OQ-45 and the ORS it was around 54–56% (see Table 7). Indeed, the χ^2 examination of the association between the measure used and the status of the client (that started the treatment in the clinical range) at the end of the treatment suggested ($\chi^2 = 20.26$; $p = .016$) that there is a significant association between the measures and the status of the client.

Finally, we wanted to establish if there was any match in identifying recovered, improved, not changed, and deteriorated participants between the measures used for evaluating the pre-post change. As can be seen in Table 8, while the match was highly statistically significant, the κ that was used as an indicator of the level of the match between the measures suggested only a ‘good match’ between CORE-OM and SCL-10R, and ‘moderate match’ between CORE-OM and OQ-45, and OQ-45 and SCL-10R. Between ORS and all others measures there was only a ‘fair match’ (We used interpretation of Landis and Koch (1977), in which the $\kappa < .20$ is Poor; .21–.40 is Fair; .41–.60 is Moderate; .61–.80 is Good; .81–1.00 is Very good). When expressed in percentages at the end of treatment, the CORE-OM matched 71% of clients to the same categories as the OQ-45, 60% to the ORS, and 78% to the SCL-10R. At the end of treatment, the OQ-45 matched 68% of participants to the same

Table 6. Categorisation of participants on the basis of reliable and clinically significant change ($N = 140$).

	No change	Improved	Recovered	Deteriorated
CORE-OM	66 (47.1%)	22 (15.7%)	48 (34.3%)	4 (2.9%)
OQ-45	85 (60.7%)	18 (12.9%)	34 (24.3%)	3 (2.1%)
ORS	94 (67.1%)	9 (6.4%)	36 (25.7%)	1 (.7%)
SCL-10R	67 (47.9%)	21 (15%)	48 (34.3%)	4 (2.9%)

Notes: CORE-OM – Clinical Outcome in Routine Evaluation – Outcome Measure; OQ-45 – Outcome Questionnaire – 45; ORS – Outcome Rating Scale; SCL-10R – Symptom Checklist – 10R.

Table 7. Categorisation of participants on the basis of reliable and clinically significant change for the participants that started in the clinical range.

	No change	Improved	Recovered	Deteriorated
CORE-OM (*n* = 107)	40 (37.4%)	22 (20.6%)	45 (42.1%)	0 (0%)
OQ-45 (*n* = 109)	59 (54.1%)	18 (16.5%)	32 (29.4%)	0 (0%)
ORS (*n* = 99)	55 (55.6%)	9 (9.1%)	35 (35.4%)	0 (0%)
SCL-10R (*n* = 101)	35 (34.7%)	21 (20.8%)	44 (43.6%)	1 (1.0%)

Notes: CORE-OM – Clinical Outcome in Routine Evaluation – Outcome Measure; OQ-45 – Outcome Questionnaire – 45; ORS – Outcome Rating Scale; SCL-10R – Symptom Checklist – 10R.

categories as the ORS, and 68% as the SCL-10R. At the end of treatment, the ORS matched 60% of clients to the same categories as theSCL-10R.

Discussion

The aim of this study was to examine three measures commonly used in assessing outcomes in routine practice (CORE-OM, OQ-45, and ORS) and one benchmark measure (SCL-10-R) as to their concordance or differences in outcome classification of therapeutic change as expressed in their pre-post effects sizes. Additionally, we wanted to examine the frequencies of the clients identified at the end of treatment as recovered, improved, not changed, and deteriorated. Furthermore, we were interested in exploring the match between these measures when classifying clients as recovered, improved, not changed, and deteriorated. Since the measures were developed as measures of therapeutic change on the basis of similar empirical research and conceptual thinking (e.g. Howard et al., 1993), we did not expect differences in their classification of therapeutic change of pre-post outcome as expressed in the mean differences (effect size) as well as in the number of recovered and improved clients at the end of treatment. We also expected that there would be a strong match between the measures in identifying the same clients as recovered, improved, not changed, and deteriorated. Our findings, however, did not correspond with those expectations.

We were surprised to discover that the CORE-OM and the SCL-10R appeared to be more sensitive to change as indicated by the number of recovered and improved clients. We were also surprised to see the pre-post differences as expressed in the effect size, with the CORE-OM showing the highest pre-post difference (pre-post effect size .98), followed by the ORS (.87), the SCL-10R (.83) and finally with the OQ-45 showing the lowest difference (.69). A sensitivity to change is the major issue for instruments measuring the outcome of therapy (cf. Burlingame, Lambert, Reisinger, Neff, & Mosier, 1995; Lambert & Hawkins, 2004; Newman, Ciarlo, & Carpenter, 1999; Sederer, Dickey, & Eisen, 1997), therefore, it is important to note that we have observed these differences. The observation, if repeated on other samples, may have relevance for the broader psychotherapy research arena, along with our thinking regarding outcome assessment in routine practice. It is crucial for

Table 8. The match between the pairs of instruments in the categorising of participants on the basis of reliable and clinically significant change ($N = 140$).

		OQ-45				
		No change	Improved	Recovered	Deteriorated	Overall
CORE-OM	No change	58 (41.4%)	3 (2.1%)	4 (2.9%)	1 (.7%)	66 (47.1%)
	Improved	10 (7.1%)	11 (7.9%)	1 (.7%)	0 (.0%)	22 (15.7%)
	Recovered	15 (10.7%)	4 (2.9%)	29 (20.7%)	0 (.0%)	48 (34.3%)
	Deteriorated	2 (1.4%)	0 (.0%)	0 (.0%)	2 (1.4%)	4 (2.9%)
	Overall	85 (60.47%)	18 (12.9%)	34 (24.3%)	3 (2.1%)	140 (100%)
$\kappa = .531$; sig. = .000; $p_0 = .71$						
		OQ-45				
		No change	Improved	Recovered	Deteriorated	Overall
ORS	No change	70 (50.0%)	10 (7.1%)	11 (7.9%)	3 (2.1%)	94 (67.1%)
	Improved	5 (3.6%)	3 (2.1%)	1 (.7%)	0 (.0%)	9 (6.4%)
	Recovered	9 (6.4%)	5 (3.6%)	22 (15.7%)	0 (.0%)	36 (25.7%)
	Deteriorated	1 (.7%)	0 (.0%)	0 (.0%)	0 (.0%)	1 (.7%)
	Overall	85 (60.7%)	18 (12.9%)	34 (24.3%)	3 (2.1%)	140 (100%)
$\kappa = .384$; sig. = .000; $p_0 = .68$						
		CORE-OM				
		No change	Improved	Recovered	Deteriorated	Overall
ORS	No change	55 (39.3%)	17 (12.1%)	18 (12.9%)	4 (2.9%)	94 (67.1%)
	Improved	4 (2.9%)	2 (1.4%)	3 (2.1%)	0 (.0%)	9 (6.4%)
	Recovered	6 (4.3%)	3 (2.1%)	27 (19.3%)	0 (.0%)	36 (25.7%)
	Deteriorated	1 (.7%)	0 (.0%)	0 (.0%)	0 (.0%)	1 (.7%)
	Overall	66 (47.1%)	22 (15.7%)	48 (34.3%)	4 (2.9%)	140 (100%)
$\kappa = .316$; sig. = .000; $p_0 = .60$						
		SCL-10R				
		No change	Improved	Recovered	Deteriorated	Overall
ORS	No change	57 (40.7%)	12 (8.6%)	21 (15.0%)	4 (2.9%)	94 (67.1%)
	Improved	3 (2.1%)	3 (2.1%)	3 (2.1%)	0 (.0%)	9 (6.4%)
	Recovered	6 (4.3%)	6 (4.3%)	24 (17.1%)	0 (.0%)	36 (25.7%)
	Deteriorated	1 (.7%)	0 (.0%)	0 (.0%)	0 (.0%)	1 (.7%)
	Overall	67 (47.9%)	21 (15.0%)	48 (34.3%)	4 (2.9%)	140 (100%)
$\kappa = .311$; sig. = .000; $p_0 = .60$						
		SCL-10R				
		No change	Improved	Recovered	Deteriorated	Overall
OQ-45	No change	57 (40.7%)	11 (7.9%)	15 (10.7%)	2 (1.4%)	85 (60.7%)
	Improved	3 (2.1%)	9 (6.4%)	6 (4.3%)	0 (.0%)	18 (12.9%)
	Recovered	6 (4.3%)	1 (.7%)	27 (19.3%)	0 (.0%)	34 (24.3%)
	Deteriorated	1 (.7%)	0 (.0%)	0 (.0%)	2 (1.4%)	3 (2.1%)
	Overall	67 (47.9%)	21 (15.0%)	48 (34.3%)	4 (2.9%)	140 (100.0%)
$\kappa = .470$; sig. = .000; $p_0 = .68$						

(*Continued*)

Table 8. (*Continued*).

		OQ-45				
		No change	Improved	Recovered	Deteriorated	Overall
		CORE-OM				
		No change	Improved	Recovered	Deteriorated	Overall
SCL-10R	No change	55 (39.3%)	3 (2.1%)	7 (5.0%)	2 (1.4%)	67 (47.9%)
	Improved	4 (2.9%)	14 (10.0%)	3 (2.1%)	0 (.0%)	21 (15.0%)
	Recovered	6 (4.3%)	4 (2.9%)	38 (27.1%)	0 (.0%)	48 (34.3%)
	Deteriorated	1 (.7%)	1 (.7%)	0 (.0%)	2 (1.4%)	4 (2.9%)
	Overall	66 (47.1%)	22 (15.7%)	48 (34.3%)	4 (2.9%)	140 (100%)
$\kappa = .650$; sig. = .000; $p_0 = .78$						

Notes: CORE-OM – Clinical Outcome in Routine Evaluation – Outcome Measure; OQ-45 – Outcome Questionnaire – 45; ORS – Outcome Rating Scale; SCL-10R – Symptom Checklist – 10R.

researchers, as well as practitioners, to select a measure that will be sensitive to change. It is important that any differences in the client's state at the end of treatment that could be attributed to psychotherapy would be captured. It seems that the observation stated by Lambert and Hill (1994), that even instruments measuring similar constructs have a different sensitivity to change, is also applicable to the instruments purposefully focusing on the areas of the clients' functioning amenable to change by psychotherapy (e.g. well-being, psychopathological symptoms, interpersonal and social functioning; Howard et al., 1993).

Interestingly, the ORS appeared to be more sensitive to change when expressed in the average pre-post effect size per the whole group, while when the classification of therapeutic change was assessed per the number of clients who recovered or improved, the instrument appeared to be less sensitive to change than the CORE-OM or the SCL-10R. This points to the fact that it is not only the particular instrument that is crucial for measuring the therapeutic change, but also how the change is measured. It seems that some measures (such as the ORS) may show large pre-post differences, but because they have low test–retest reliability and therefore a larger RCI, when we are assessing an individual's pre-post change (recovered, improved, not changed, deteriorated) the instrument is more conservative and so returns a higher number of the clients as 'not changed'.

What could explain the differences between the instruments? The obvious candidates are the actual wording of the instruments and their content. Although the content areas of the instruments are similar, the OQ-45 also includes items focusing on drug and alcohol abuse, which are not mentioned in the CORE-OM or SCL-10R. As these may not be pertinent for the whole sample, this may decrease the variability for the whole instrument. Also the OQ-45 (and to a certain degree the ORS) has a larger proportion of items focusing on social roles and the settings in which it is applied (work, school, home) and interpersonal relationships (it has two domains focused on this areas while the CORE-OM has just one). The area of interpersonal relationships together with the social roles changes later on in therapy (Howard et al., 1993), so given that the OQ-45 has a higher proportion of items in this area, this may explain the lower pre-post change on the OQ-45.

Given that we found these differences among the instruments in terms of their sensitivity to pre-post change, it is not surprising that the match between which clients the instruments identified as recovered, improved, not changed, and/or deteriorated was only moderate. What is perhaps unexpected is the relatively low match between the OQ-45 and the ORS, given that the ORS was developed as an ultra-brief visual analogue scale based on the OQ-45. It is interesting that the match is quite small, despite the relatively high correlations between the instruments (for instance for the OQ-45 and ORS it was .59 (Miller et al., 2003) or in the current sample .65). The ORS had a particularly low match with each of the other instruments; it seems that the use of a visual-analogue scale asking for an expression of the amount of positive sense of self-assessed functioning (ORS), may capture different aspects of the client's mental health status than the traditional self-report measures which focus on problematic functioning.

All of these observations suggest that an assumption of a straightforward generalisation and comparison from one study to another is problematic. Even very similar instruments developed on the basis of similar theoretical conceptualisations and empirical findings, may be obtaining different pre-post outcomes. This finding is in a sense somewhat disturbing. It has particular implications for meta-analyses and systematic reviews, which often decontextualise the effect sizes or recovery (improvement) rates that they deal with. Meta-analyses look sophisticated and robust and thus they often silence the criticism that they may be mixing apples and oranges. We would therefore very much caution against the de-contextualisation of the psychotherapy outcome assessment as clearly the use of different instruments may have significant implications.

Limitations

The main advantage of our project was the unique within-subjects design, as all four measures were administered to the same sample of participants and thus allowed their direct comparison for outcome classification of therapeutic change. The fact that we calculated the cut-off scores and the RCIs on the basis of the same samples of participants is also the strength of the design. This is probably the first study that examined concordance or differences in outcome classification of therapeutic change in commonly used instruments which measure the outcome in routine practice, using the cut-off scores and the RCIs developed on the basis of the same clinical and non-clinical samples, as well as test–retest data.

There are, however, some obvious limitations to our study. For instance, we used the Slovak versions of the instruments, so generalisability to other language mutations of those instruments, or indeed to the original instruments, has to be done with caution. Probably the biggest problem was our sample size. Both the clinical and the non-clinical samples were quite small.

Furthermore, the clinical sample was dominated by participants from psychiatric inpatient facilities. This may have resulted in this sample containing clients with a more severe psychopathology and thus making it less representative of the entire clinical population. Furthermore, it is possible that the CORE-OM and the SCL-10R is more suitable for this type of participant as the OQ-45 (and the ORS on which it was based) was mainly developed with a student counselling population (although authors see as appropriate also for inpatient population; cf. Lambert & Finch, 1999).

In terms of the clinical population, a further limitation is the variation regarding the length of treatment (4 days to 41 weeks). It can be expected that a shorter length of treatment will only be able to detect smaller changes. This has implications for the detection of differences in pre-post change among the studied instruments. Therefore, we would recommend that future studies use a sufficient length of treatment and perhaps that the sample that would undergo the same length of treatment.

Furthermore, we do not know 'the real' change, therefore, we cannot truly speak about the sensitivity to change, but rather about the magnitude of change and the number of recovered and improved clients as measured by one instrument in relative comparison to the other. Thus none of the instruments can be identified as superior to the other in terms of identifying the real change. In any case, we eagerly await further studies examining these and other measures, in order to learn more about the methodological issues that arise when using different outcome measures, to assess psychotherapy outcome.

Notes on contributors

Matus Biescad, PhD, graduated with a doctoral degree in psychology from Trnava University, Slovakia. His dissertation addressed outcome assessment in routine practice. He completed his psychotherapeutic training in humanistic psychotherapy and worked as a psychologist at a psychiatric hospital.

Ladislav Timulak, PhD, is associate professor at Trinity College Dublin. He is course director of the doctorate in counselling psychology. He is involved in the training of counselling psychologists and various psychotherapy trainings as well. He is both an academic and a practitioner.

References

Barkham, M., Evans, C., Margison, F., McGrath, G., Mellor-Clark, J., Milne, D., & Connell, J. (1998). The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Journal of Mental Health, 7*, 35–47.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., … 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*, 184–196.

Barkham, M., & Mellor-Clark, J. (2003). Bridging evidence-based practice and practice-based evidence: Developing a rigorous and relevant knowledge for the psychological therapies. *Clinical Psychology and Psychotherapy, 10*, 319–327.

Barkham, M., Mellor-Clark, J., Connell, J., & Cahill, J. (2006). A core approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System. *Counselling and Psychotherapy Research, 6*, 3–15.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56*, 893–897.

Bieščad, M. (2004). *Psychometrická analýza sebaposudzovacej škály Symptom Checklist – 90 (SCL-90)* [Psychometric analysis of the Symptom Checklist – 90, a self-report] (Diplomová práca [Master's dissertation]). Trnava: Fakulta humanistiky, Trnavskej univerzity.

Bieščad, M. (2007). *Aplikácia nástrojov merajúcich výsledky psychoterapie. Porovnanie citlivosti nástrojov merania v jednotlivých oblastiach terapeutickej zmeny* [The application of psychotherapy outcome instruments: A comparison of the instruments' sensitivity in different areas of therapeutic change] (Dizertačná práca [PhD dissertation]). Trnava: Filozofická fakulta, Trnavskej univerzity v Trnave.

Biescad, M., & Timulak, L. (2006, June 21–24). *Using CORE-OM, OQ-45 and Outcome Rating Scale in various mental health care settings: Preliminary data comparing results on the three instruments*. Paper presented at 37th annual meeting of the Society for Psychotherapy Research, Edinburgh, Scotland.

Boleloucký, Z. (1993). Psychiatrické posuzovací stupnice a dotazníkové metódy [Psychiatric rating scales and questionnaires]. In J. Baštecký, J. Šavlík, & J. Šimek (Eds.), *Psychosomatická medicína* [Psychosomatic medicine] (s. 118–127). Praha: Grada Avicenum.

Burlingame, G. M., Lambert, M. J., Reisinger, C. W., Neff, W. M., & Mosier, J. (1995). Pragmatics of tracking mental health outcomes in a managed care setting. *The Journal of Mental Health Administration, 22*, 226–236.

Cahill, J., Barkham, M., Stiles, W. B., Twigg, E., Hardy, G. E., Rees, A., & Evans, C. (2006). Convergent validity of the CORE measures with measures of depression for clients in cognitive therapy for depression. *Journal of Counseling Psychology, 53*, 253–259.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cone, J. D. (2001). *Evaluating outcomes: Empirical tools for effective practice*. Washington, DC: American Psychological Association.

Connell, J., & Barkham, M. (2007). *CORE-10 user manual, Version 1.1* (pp. 1–40). CORE System Trust & CORE Information Management Systems Ltd.

CORE System Group. (1998). *CORE System (information management) handbook*. Leeds: Author.

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

Derogatis, L. R., Lipman, R. S., & Covi, L. (1973). SCL-90: An outpatient psychiatric rating scale-preliminary report. *Psychopharmacology Bulletin, 9*, 13–28.

Doerfler, L. A., Addis, M. E., & Moran, P. W. (2002). Evaluating mental health outcomes in an inpatient setting: Convergent and divergent validity of the OQ-45 and BASIS-32. *Journal of Behavioral Health Services and Research, 29*, 394–403.

Elfström, M. L., Evans, C., Lundgren, J., Johansson, B., Hakeberg, M., & Carlsson, S. G. (2012). Validation of the Swedish version of the Clinical Outcomes in Routine

Evaluation Outcome Measure (CORE-OM). *Clinical Psychology and Psychotherapy, 20*, 447–455.

Evans, C. (2012a). Cautionary notes on power steering for psychotherapy. *Psychologie Canadienne* [Canadian Psychology], *53*, 131–139.

Evans, C. (2012b). The CORE-OM (Clinical Outcomes in Routine Evaluation – Outcome Measure) and its derivatives. *Integrating Science and Practice, 2*, 12–15.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Toward a standardised brief outcome measure: Psychometric properties and utility of CORE-OM. *British Journal of Psychiatry, 180*, 56–60.

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical Outcomes in Routine Evaluation. *Journal of Mental Health, 9*, 247–255.

Gampe, K., Bieščad, M., Balúnová-Labaničová, L., Timuľák, L., & Evans, Ch. (2007). Slovenská adaptácia metódy CORE-OM [Slovak adaptation of the CORE-OM]. *Česká a Slovenská psychiatrie* [Czech and Slovak Psychiatry], *103*, 4–13.

Holi, M. (2003). *Assessment of psychiatric symptoms using the SCL – 90*. Helsinky: Helsinky University Printing House.

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of Consulting and Clinical Psychology, 61*, 678–685.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51*, 1059–1064.

International Centre for Clinical Excellence. (2013). *The Outcome Rating Scale*. Retrieved from http://www.centerforclinicalexcellence.com/site.php?page=measures.php

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*, 12–19.

Klaghofer, R., & Brähler, E. (2001). Konstruktion und teststatistische prüfung einer kurzform der SCL-90-R [Construction and statistical evaluation of a short version of the SCL-90–R]. *Zeitschrift für Klinische Psychologie, Psychiatrie und Psychoterapie* [Journal of Clinical Psychology, Psychiatry and Psychotherapy], *49*, 115–124.

Lambert, M. J. (1983). Introduction to assessment of psychotherapy outcome: Historical perspective and current issues. In M. J. Lambert, E. R. Christensen, & S. S. DeJulio (Eds.), *The assessment of psychotherapy outcome* (pp. 3–32). New York, NY: Wiley.

Lambert, M. J. (2010). *Prevention of treatment failure*. Washington, DC: American Psychological Association.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy, 3*, 249–258.

Lambert, M. J., & Finch, A. E. (1999). The Outcome Questionnaire. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 831–869). Mahwah, NJ: Lawrence Erlbaum Associates.

Lambert, M. J., Hannöver, W., Nisslmüller, K., Richard, M., & Kordy, H. (2002). Fragebogen zum Ergebnis von Psychotherapie: Zur Reliabilität und Validität der deutschen Übersetzung des Outcome Questionnaire 45.2 (OQ-45.2) [Psychotherapy outcome measures: Toward reliability and validity of the German version of the Outcome Questionnaire 45.2 (OQ-45.2)]. *Zeitschrift für Klinische Psychologie und Psychotherapie* [Journal of Clinical Psychology, Psychiatry and Psychotherapy], *31*, 40–46.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*, 159–172.

Lambert, M. J., & Hawkins, E. J. (2004). Measuring outcome in professional practice: Considerations in selecting and using brief outcome instruments. *Professional Psychology: Research and Practice, 35*, 492–499.

Lambert, M. J., & Hill, C. E. (1994). Assessing psychotherapy outcomes and processes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 72–113). New York, NY: Wiley.

Lambert, M. J., Morton, J. J., Hattfield, D., Harmon, C., Hamilton, S., Reid, R. C., … Burlingame, G. M. (2004). *Administration and scoring manual for the OQ© – 45.2 (Outcome Questionnaire)*. Orem, UT: American Professional Credentialing Services.

Lambert, M. J., Okiishi, J. C., Finch, A., & Johnson, L. D. (1998). Outcome assessment: From conceptualization to implementation. *Professional Psychology: Research and Practice, 29*, 63–70.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice, 10*, 288–301.

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11*, 49–68.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lo Coco, G., Chiappelli, M., Prestano, C., Gullo, S., Bensi, L., Lo Verso, G., & Lambert, M. J. (2006, June 21–24). *Generalizability of normative data for the OQ-45: A study with an Italian outpatient group*. Poster presented at 37th annual meeting of the Society for Psychotherapy Research, Edinburgh, Scotland.

Machado, P. P. P., & Klein, J. M. (2006, June 21–24). *The Outcome Questionnaire – 45: Portuguese psychometric data with a non-clinical sample*. Poster presented at 37th annual meeting of the Society for Psychotherapy Research, Edinburgh, Scotland.

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*, 91–100.

Miller, S. D., Duncan, B. L., & Hubble, M. A. (2004). Beyond integration: The triumph of outcome over process in clinical practice. *Psychotherapy in Australia, 10*, 2–19.

Miller, S. D., Duncan, B. L., & Hubble, M. A. (2005). Outcome-informed clinical work. In J. C. Norcross & M. R. Goldfried (Eds.), *Handbook of psychotherapy integration* (2nd ed., pp. 84–102). New York, NY: Oxford University Press.

Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome management system. *Journal of Clinical Psychology, 61*, 199–208.

Newman, F. L., Ciarlo, J. A., & Carpenter, D. (1999). Guidelines for selecting psychological instruments for treatment planning and outcome assessment. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 153–170). Mahwah, NJ: Lawrence Erlbaum Associates.

Nguyen, T. D., Attkisson, C. C., & Stegner, B. L. (1983). Assessment of patient satisfaction: Development and refinement of a Service Evaluation Questionnaire. *Evaluation and Program Planning, 6*, 299–313.

Ogles, B. M., Lambert, M. J., & Fields, S. A. (2002). *Essentials of outcome assessment*. New York, NY: Wiley.

Palmieri, G., Evans, C., Hansen, V., Brancaleoni, G., Ferrari, S., Porcelli, P., … Rigatelli, M. (2009). Validation of the Italian version of the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM). *Clinical Psychology and Psychotherapy, 16*, 444–449.

Rosen, C. S., Drescher, K. D., Moos, R. H., Finney, J. W., Murphy, R. T., & Gusman, F. (2000). Six and ten-item indexes of psychological distress based on the Symptom Checklist-90. *Assessment, 7*, 103–111.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Sederer, L. I., Dickey, B., & Eisen, S. V. (1997). Assessing outcomes in clinical practice. *Psychiatric Quarterly, 68*, 311–325.

Skre, I., Friborg, O., Elgarøy, S., Evans, C., Myklebust, L. H., Lillevoll, K., … Hansen, V. (2013). The factor structure and psychometric properties of the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM) in Norwegian clinical and non-clinical samples. *BMC Psychiatry, 13*, 99.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74*, 242–261.

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Is it sensitive to changes in counseling center clients? *Journal of Counseling Psychology, 51*, 38–49.

Wiger, D. E., & Solberg, K. B. (2001). *Tracking mental health outcomes*. New York, NY: Wiley.

Zvarová, M., & Bieščad, M. (2007, April). *Overenie konštruktovej validity a vzájomných vzťahov Dotazníka výsledku (Outcome Questionnaire – 45) a Škály hodnotenia výsledku (Outcome Rating Scale).* [Evaluation of construct validity and of the relationship between the Outcome Questionnaire 45 and Outcome Rating Scale]. In *II. Medzinárodnej konferencii doktorandov odborov Psychológia a Sociálna práca* [2nd International conference of doctoral students in psychology and social work], Nitra.