Received: 15 June 2018 | Revised: 7 January 2019 | Accepted: 12 January 2019

DOI: 10.1002/mpr.1774

ORIGINAL ARTICLE

WILEY

Comparing outcomes: The Clinical Outcome in Routine Evaluation from an international point of view

Marina Zeldovich | Rainer W. Alexandrowicz

Institute of Psychology, Alpen-Adria Universität Klagenfurt, Klagenfurt, Austria

Correspondence
Marina Zeldovich, Institute of Psychology, Alpen-Adria Universität Klagenfurt, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria.
Email: marina.zeldovich@aau.at

Abstract

Objectives: The Clinical Outcome in Routine Evaluation–Outcome Measure (CORE-OM) is a freely accessible self-assessment questionnaire with a total of 34 items measuring the progress of psychological or psychotherapeutic treatments according to four scales (well-being, problems, functioning, and risk). The CORE-OM originated in the United Kingdom and has been translated into 54 languages and dialects. The aim of this study is to systematically compare the translated versions.

Method: A total of 21 translations were compared using methods of systematic review and meta-analysis.

Results: The results show a certain heterogeneity between the studies; however, the 21 translations can be declared as equivalent.

Conclusion: The factorial structure could not be replicated in any of translations. Therefore, further analysis of the CORE-OM domains is recommended. In addition, some supplementary restrictions on the translation process, data collection, and reporting of results are necessary to ensure comparability and quality of CORE-OM translations.

KEYWORDS

CORE-OM, meta-analysis, outcome measurement, translation of questionnaires

1 | INTRODUCTION

Outcome measurement (OM) is the assessment of the "effect on a patient's health status that is attributable to an intervention by a health professional or health service" (Andrews, Peters, & Teesson, 1994, p. 3). An outcome is the end result of a provided health service that affects the health status and functioning of the patient treated, thus reflecting what happens to the patient "in terms of palliation, control of illness, cure, or rehabilitation" (Brook, Williams, & Avery, 1976, p. 809). It focuses on the change and allows evaluation of both the services provided and the patients' progress during treatment. In the late 1970s, a discussion regarding the necessity of OM in mental health services arose in the United States and spread internationally in the 1980s (Brook et al., 1976; Erickson, 1975; Lohr, 1988; Slade, Thornicroft, & Glover, 1999). Sutherland and Till (1993) identified three levels, at which OM might prove useful: micro (e.g., individual or clinical), meso (e.g., within an institution), and macro (e.g., governmental).

1.1 | OM instruments

OM instruments assess outcomes using either an external perspective (e.g., clinicians or relatives) or the patient's subjective perception, or

both (Thornicroft & Slade, 2014). They cover various domains, such as well-being, global functioning, quality of life, physical or mental condition, satisfaction with the treatment, provided services, or their costs (Slade, 2002; Thornicroft & Tansella, 2010). Outcome measures can be symptom independent or targeted a specific group of mental diseases; they can focus on recovery or individual goals in the course of treatment (Thornicroft & Slade, 2014).

Several instruments have been developed to measure outcomes. For example, the Health of the Nation Outcome Scales (Wing et al., 1998) focuses upon the outcome of mental health treatments using an external perspective (clinicians). Because the external perspective is not always reliable and possibly biased, some authors recommend to integrate also the patient's perspective (Slade, 1996; Slade, Leese, Taylor, & Thornicroft, 1999). The Inventory of Interpersonal Problems (Horowitz, Rosenberg, Baer, Ureno, & Villasenor, 1988) focuses on interpersonal problems causing psychological distress. Symptom dependent measures such as the Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996) can be used for assessment of the symptom severity in course of the therapies. Moreover, there are questionnaires measuring important constructs for outcome assessment, such as well-being (e.g., Quality of Well-Being Scale; Kaplan & Bush, 1982), global functioning (e.g., Global Assessment Scale; Endicott, Spitzer, Fleiss, & Cohen, 1976, or Work and Social Adjustment Scale; Mundt, Marks, Shear, & Greist, 2002), and quality of life (e.g., Quality of Life Scale; Flanagan, 1978).

1.2 | The CORE-OM

The present study focuses on the Clinical Outcome in Routine Evaluation–Outcome Measure (CORE-OM; Barkham et al., 1998). It concentrates on the areas mentioned above plus the patient's resources. If the CORE-OM works the way it was designed, we would dispose of an instrument gathering a wide spectrum of information useful for assessing the treatment progress from the patient's perspective. Its development relied on a survey of mental health services in the United Kingdom, which revealed a lack of systematic information on patient's health status in pretreatment and posttreatment phases (Mellor-Clark, Barkham, Connell, & Evans, 1999). The CORE-OM has been integrated into the British National Health System (Slade, 2010).

1.3 | Structure

The CORE-OM contains 34 items split into to the four scales *well-being* (four items), *problems/symptoms* (12 items), *functioning* (12 items), and *risk* (six items) using a five-categorical response format (0 = *not at all*, 1 = *only occasionally*, 2 = *sometimes*, 3 = *often*, and 4 = *most or all the time*); eight items are inversely worded. According to the manual, the CORE-OM is not restricted to specific diagnosis groups. The questionnaire is copyleft (i.e., it can be used free of charge), which fosters its broad application.

1.4 | Evaluation

Evans et al. (2002) evaluated the psychometric properties of the CORE-OM using a clinical sample collected from 23 sites within the National Health Service, three university student counselling services, and a staff support service in the United Kingdom (n = 890; 60% female) and a nonclinical sample (n = 1,106; 54% female) of university students and staff as well as "a sample of convenience" (Evans et al., 2002, p. 53; Lohr, 1999) representing the "general population." Table 1 gives an overview about the results of psychometric analyses.

1.5 | Translation

To foster international comparisons, Evans, Mellor-Clark, Marginson, and Barkham (2000) and Evans et al. (2002) called for translations into other languages by psychologists, psychiatrists, or psychotherapists. This process has to meet specific requirements defined by CORE System Trust (CST, 2011, 2015), requiring (a) forward translation, (b) a focus group discussing the translation, and (c) field testing and backward translation. One of the authors of the English CORE-OM has to accompany the translational procedure. The CST (2011, 2015) further specifies rules regarding the examination of the psychometric

TABLE 1 An overview of psychometric properties (reliability and validity) of the English CORE-OM in accordance to Evans et al. (2002)

	Method/ instrument	Sample type	*N*	Well-being 4 items	Problems 12 items	Functioning 12 items	Risk 6 items	Nonrisk items 28 items	Total score 34 items
Reliability	Cronbach's α	Clinical	1,106	0.75	0.88	0.87	0.79	0.94	0.94
		Nonclinical	890	0.77	0.90	0.86	0.79	0.94	0.94
	Test stability (ρ)	Student sample	43	0.88	0.87	0.87	0.64	0.91	0.90
Validity	BDI-I	Clinical	251	0.77	0.78	0.78	0.59	0.84	0.85
	BDI-II		29	0.79	0.74	0.78	0.32	0.83	0.81
	BSI		97	0.63	0.76	0.71	0.62	0.79	0.81
	SCL-90		34	0.68	0.87	0.79	0.83	0.85	0.88
	GHQ		69	0.67	0.66	0.65	0.56	0.72	0.75
	GHQ-A		69	0.43	0.60	0.44	0.30	0.56	0.55
	GHQ-B		69	0.55	0.61	0.57	0.30	0.64	0.64
	GHQ-C		69	0.60	0.52	0.60	0.44	0.62	0.63
	IIP-32		246	0.48	0.58	0.65	0.45	0.64	0.65

Note. CORE-OM: Clinical Outcome in Routine Evaluation–Outcome Measure; BDI-I: Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961); BDI-II: Beck Depression Inventory (Beck et al., 1996); BSI: Brief Symptom Inventory (Derogatis & Melisaratos, 1983); GHQ: General Health Questionnaire, 28-item version (Goldberg & Hiller, 1979); GHQ-A: somatic symptoms; GHQ-B: anxiety and insomnia; GHQ-C: social dysfunction; IIP-32: Inventory of Interpersonal Problems, 32-item version (Barkham, Gillian, & Startup, 1996); SCL-90: Symptom Checklist-90-Revised (Derogatis, 1983).

properties of the translations. The sample should be representative of the target population and comprise at least $N = 100$ for a clinical population, $N = 40$ for test–retest examination with the interval of 1 week to 1 month, and $N = 200$ for the clinical and $N = 200$ for the nonclinical population. Given that internal consistency and retest reliability can be considered sufficient, the CST guidelines further recommend assessing both the reliable change (Jacobson, Follette, & Revenstorf, 1984) and clinical significant change (CSC; Jacobson & Truax, 1991), the latter also termed "clinical reliability" (Evans, Margison, & Barkham, 1998). The CST (2011) further recommends exploratory factor analysis (EFA) of the data. Interestingly, the CST (2015) relaxed the requirements by recommending sweepingly $N = 100$ for all samples and dropping the factor analysis from the list.

Barkham, Mellor-Clark, Connell, and Cahill (2006) emphasized that the CORE-OM focuses on "ethnic groups and European languages" (p. 9). Since its introduction in 1999, the CORE-OM has been translated into 52 languages and dialects. According to the CST (2018), 23 translated versions of the CORE-OM have been published and 29 translations were under development at the time of writing.

1.6 | Rationale

According to the World Health Organization (WHO, 2018) guidelines, a translation should be "conceptually equivalent in each of the target countries/cultures. (…) should be equally natural and acceptable and should practically perform in the same way. (…) A well-established method to achieve this goal is to use forward-translations and back-translations." Thus, the WHO guidelines consider forward and backward translation sufficient to achieve "conceptual equality" of questionnaire translations. The translation process of the CORE-OM followed these steps, so that from a WHO perspective, the translated versions of the CORE-OM could be considered equivalent. However, if that claim holds, we should be able to provide empirical evidence for it—or reveal a lack thereof.

1.7 | Objectives

The authors of the present study are not aware of a systematic comparison of the published translations of the CORE-OM so far. Goldhahn, Shisha, Macdermid, and Goldhahn (2013) emphasized the importance of "appropriately translated instruments" for "international multicentre studies; or inclusion of people with different cultural backgrounds in national trials" (p. 591). Because the CORE-OM will be used for international comparisons, we require empirical evidence that the translations can be considered equivalent from a psychometric point of view. Therefore, the present study compares all available studies of the translations of the CORE-OM with respect to the reported psychometric properties considering the three major criteria (a) reliability (as reflected by internal consistency and retest stability), (b) validity (in terms of factorial structure, convergent, and discriminant validity), and (c) objectivity (in terms of application, evaluation, and interpretation of results). Only if these criteria are met to a sufficient extent can we recommend the various CORE-OM versions for comparisons across countries with differing languages.

2 | METHOD

2.1 | Study design

The present study uses techniques applied in systematic reviews and meta-analyses. Each article presenting a translation of the CORE-OM serves as a primary study, the results of which will be summarised. We follow the guidelines for Preferred Reporting Items for Systematic Reviews and Meta-Analysis (Moher et al., 2015). First, we provide a systematic review of studies of CORE-OM translations with respect to the translational processes and psychometric analyses performed therein. Second, we conduct a meta-analysis on the psychometric details with the primary focus on the three major criteria reliability, validity, and objectivity.

2.2 | Data collection

The official website of the CST (2018) lists available translations, contact information, and translations currently in progress. For each language listed there, we conducted a search on PsychINFO and PubMed using the search terms CORE-OM AND psychom* propert* OR CORE-OM AND translat*; publication year from 1998 to 2018. Moreover, we also performed a Google scholar search using generic search terms, that is, "CORE-OM" along with the respective language, for example, "German CORE-OM". Authors, whose translations could not be found in these sources, were contacted via e-mail. The target was to collect all articles presenting a translation of the CORE-OM. To ensure the validity of findings, two persons (M. Z. and L. C. W.) performed the search independently of each other.

2.3 | Information extraction

From each article, we extracted (a) data collection and sampling characteristics (sample type, recruiting of participants, and duration); (b) descriptive statistics of clinical and nonclinical samples (sample size, gender, and mean age); (c) reliability measures (Cronbach's α); test stability coefficients (Spearman and Pearson); and (d) details regarding the examination of the validity (factorial structure using EFA and confirmatory factor analysis [CFA] and correlations between the CORE-OM and the SCL-90 and the BDI-II).

2.4 | Analysis

Using a random-effect meta-analytical approach (DerSimonian & Laird, 1986), we pooled Cronbach's α, the stability coefficients, and the correlations of the total scores of SCL-90 and BDI-II with the CORE-OM total score, the modified total score (nonrisk items), and the scale scores. We calculated Cochran's Q (Cochran, 1950) and I^2 and H^2 (Higgins & Thompson, 2002) to assess the variation of studies outcomes. Further, we generated the diagnostic plot of Baujat, Mah, Pignon, and Hill (2002) and forest plots (Lewis & Clarke, 2001) for visualisation of the results (see Supporting Information). Regarding I^2, we follow the guideline of Quintana (2015), who suggests to consider up to 25% as low, 50% as moderate, and 75% and above as high

TABLE 2 An overview of all identified CORE-OM publications

					Clinical								Nonclinical						
									Gender		Age					Gender		Age	
No.	Author(s), year	Translation	Pub	Lang	*N*	Patients	Type	Collection	Female	Male	*M*	*SD*	*N*	Type	Collection	Female	Male	*M*	*SD*
[1]	Evans et al. (2002)	English	Paper	En	890	Out	Mental health services	pp	530	344	—	—	1,106	University staff, students	pp	601	498	—	—
[2]	Bodinaku (2014)	Albanian	PhD	En	209	Out	Mental health centre	pp	129	80	37	12.4	501	Community	pp	249	252	40.2	15.1
-[a]	Cartasso and Lemos (2012)	Argentinian	Paper	Esp	106	Out	Psychiatric centre	pp	75	31	—	—	—	—	—	—	—	—	—
-[a]	Santana et al. (2015)	Brazilian Portuguese	Paper	En	44	Out	Trauma clinic	pp	20	24	43.2	14.1	55	Community	pp	24	31	37	17.4
[3]	Jokic-Begic, Korajlija, and Jurin (2014)	Croatian	Paper	Hr	183	In	Psychiatric hospital	pp	108	75	47	10.68	425	Community	pp	231	194	38.7	12.29
[4]	Juhová (2015)	Czech	PhD	Cz	175	In	Psychiatric hospital	pp	107	66	—	—	300	Students, community	pp/online	237	62	—	—
[5]	Meerding, van't Spijker, and van Riessen (2012)	Dutch	Paper	Nl	10,988	Out	Different mental praxes	pp	7,119	3,868	38.4	13	613	Community	pp	314	298	47.7	17.7
[6][b]	Juntunen, Piiparinen, Honkalampi, Inkinen, and Laitila (2015)	Finnish	Paper	Fin	—	—	—	—	—	—	—	—	209	Community	pp	114	95	43.2	16.2
	Honkalampi et al. (2017)		Paper	En	201	In/out	Mental health services	pp	121	79	35.02	12.76							
[7]	Sproll (2011)	German	Dipl	Ger	179	Out	Psychotherapy ambulance, primary care ambulance	pp	103	76	36.6	12.6	197	Community	pp	100	97	30.99	12.99
[8]	Kristjánsdóttir et al. (2015)	Icelandic	Paper	En	387	Out	University hospital	pp	317	70	38.05	—	207	Students	pp	136	71	22.7	—
[9]	Palmieri et al. (2009)	Italian	Paper	En	647	In/out	Psychotherapy ambulance	pp	443	196	36	11.9	263	Students	pp	192	71	25	5.8
[10]	Uji, Sakamoto, Adachi, and Kitamura (2012)	Japanese	Paper	En	1,357	In	Psychiatric hospitals	pp	881	433	35	13.9	—	Students, community	pp	—	—	—	—
[11]	Viliuniene et al. (2012)	Lithuanian	Paper	En	39	Out	Psychotherapy ambulance	pp	23	16	36.7	12.5	133	Students	pp	93	37	20.8	0.7
[12]	Skre et al. (2013)	Norwegian	Paper	En	527	Out	Psychiatric centre	pp	320	207	37.4	12.6	464	Students, community	pp	333	131	32.6	14.3
-[a]	Sales, de Matos Moleiro, Evans, and Alves (2012)	Portuguese	Paper	Port	—	—	—	—	—	—	—	—	—	Students, community	pp	77	34	14	62
[13]	Zeldovich, Ivanov, Evans, and Andreas (2014)	Russian	Paper	Ru	159	In	Psychiatric hospital	pp	79	80	38.2	12.8	224	Students, community	pp/online	169	55	26.5	8.8
[14]	Gampe, Bieščad, Balúnová-Labanicová, Timulák, and Evans (2012)	Slovak	Paper	Slo	40	In/out	Psychiatric hospital	pp	20	20	36.13	—	74	Students, community	pp	38	31	23.27	—
[15]	Feixas et al. (2012) and Trujillo et al. (2016)	Spanish	Paper	Esp, en	192	Out	Primary care ambulance	pp	130	61	41.3	14.9	452	Students, community	pp	343	94	29.3	14.4

(Continues)

TABLE 2 (Continued)

					Clinical							Nonclinical							
									Gender		Age					Gender		Age	
No.	Author(s), year	Translation	Pub	Lang	*N*	Patients	Type	Collection	Female	Male	*M*	*SD*	*N*	Type	Collection	Female	Male	*M*	*SD*
[16]	Elfström et al. (2012)	Swedish	Paper	En	619	Out	Primary care ambulance	pp	427	171	40.5	13	229	Students	pp	126	103	27.5	8
[17]	Campbell and Young (2016)	Xhosa	Paper	En	49	Out	Mental health services for students	pp	–	–	–	–	194	Students	pp	–	–	–	–
[18]	Campbell (2011)	South African English	Paper	En	312	Out	Mental health services for students	pp	232	80	20.6	–	421	Students	pp	256	165	21.23	–

Note. In the text, we refer to the numbers in the first column. CORE-OM: Clinical Outcome in Routine Evaluation–Outcome Measure; Pub: publication; Paper: published paper; PhD: PhD thesis; Dipl: diploma thesis; Lang: language of the publication. *N*: sample size; Type: sample type; pp: paper–pencil; Online: online survey; Collection: type of data collection; *M*: mean; *SD*: standard deviation; —: information not available.

[a]No psychometrics.

[b]Psychometric properties of the Finnish CORE-OM for the nonclinical sample originate from the publication Juntunen et al. (2015) and are, therefore, considered only once.

variance between the studies. Using Cook's distance (Cook & Weisberg, 1982), we identified the studies contributing most to heterogeneity. To detect possible explanations of observed differences in Cronbach's α and convergent validity coefficients, we performed a moderator analysis using (a) mean age of participants, (b) gender, and (c) sample type (inpatients, outpatients, and mixed samples) as covariates. For test stability coefficients, we disposed only of information on the sample type (community, students, and mixed samples), which was used as a moderator. Age and gender (proportion of females) were introduced as quantitative covariates, and sample type was dummy coded (*outpatients* as reference group for the internal consistency's analyses and *community* for test stability).

All analyses were performed with R (R Core Team, 2013) applying the packages robumeta (Fisher, Tipton, & Zhipeng, 2017) and metafor (Viechtbauer, 2010). For better readability, references to specific studies are given in brackets throughout the text (full list in Table 2).

3 | RESULTS

3.1 | Study selection

From the 52 translations listed on the CST (2018), we could identify 26 publications covering 21 translations in the literature search. Two versions had more than one publication each: The Spanish version had one Spanish and one English paper, and the Finnish version had one master thesis and two publications in Finnish and English, respectively. If available, we chose the most recent available article for analysis, with a preference for published works over unpublished manuscripts. That way, 19 papers could be included in the quantitative analysis. Figure 1 shows the attrition diagram of the extraction process.

Some of these 19 articles did not report all information we sought for: 18 publications reported Cronbach's α, 12 coefficients for test stability examination, and five correlation analyses of the corresponding CORE-OM with BDI-II [1, 2, 8, 14, 15] and six with SCL-90 [1, 2, 6, 7, 14, 15]. Four studies were not considered in our analyses for they used other instruments for validation (BDI-I [6], Inventory of Interpersonal Problems-32 [7], Beck Anxiety Inventory [8], and Brief Symptom Inventory [13]). Regarding dimensional analysis, we could identify five publications applying principal component analysis (PCA) and four studies provided a CFA, but results were not reported in sufficient detail. Due to the small number of studies applying EFA and CFA and the lack of comprehensive information on the fit statistics provided by those who did, we cannot perform a quantitative analysis and will, therefore, only summarise the results. For the same reason, no analysis of CSC could be performed. These findings were verified by two independent persons M. Z. and A. M. K.

3.2 | Sample characteristics

From 21 translations, the clinical samples totalled to N = 17,303, with 11,184 (65%) female and 5,977 (35%) male respondents (142 missing); the mean age was 37.3 (SD = 12.9) years. All clinical samples were

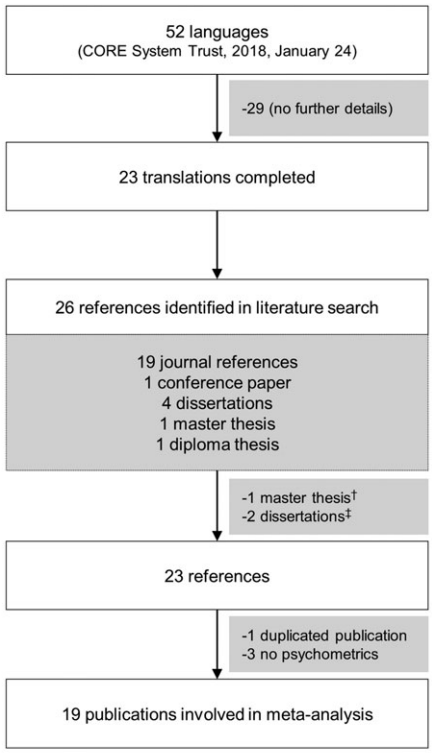

FIGURE 1 A flow diagram depicting the selection process of the included in the analysis papers

convenience samples covering both outpatients and inpatients of hospitals, primary care, day and psychotherapeutic services, private psychologists, and psychotherapists. All studies used the paper–pencil version of the respective CORE-OM translation. Nonclinical samples consisted of *n* = 3,633 (61%) female and *n* = 2,319 (39%) male respondents. Gender was missing for *n* = 257 (1%) respondents. The mean age in the nonclinical samples was 30.4 (*SD* = 15.8). Three of the 21 studies tried to adapt the proportions in nonclinical samples using sociodemographic factors from population statistics of corresponding countries, regions, or cities. Bodinaku (2014) [2] used a random walk technique for data collection in Albania, only Meerding et al. (2012) [5] engaged a survey agency for recruiting of respondents in the Netherlands. Student samples were used in 14 studies as nonclinical samples, six of which as the only source of information. All studies used the paper–pencil version, and two had additionally an online survey. Reported durations for data collection were between half a year and 2 years. Table 2 gives an overview of field phases characteristics of the studies.

3.3 | Internal consistency

In total, 18 studies calculated indices of internal consistency using the Cronbach's α for the total and subscale scores, 15 of them reported also values for the nonrisk total score (Table 3). In addition to the total score for all 34 items, Evans et al. (2002) suggested determining the total score for nonrisk items (28 items without six items from the *risk* scale) to investigate psychological distress, which will be included in

TABLE 3 Reliability of the CORE-OM translations

	Reliability															
	Cronbach's α						Test stability									
No.	W	P	F	R	T	-R	*N*	Type	Timespan	Coef	W	P	F	R	T	-R
[1]	0.75	0.88	0.87	0.79	0.94	0.94	43	Students	—	S	0.88	0.87	0.87	0.64	0.90	0.91
[2]	0.60	0.89	0.78	0.86	0.92	0.91	104	Community	7d	—	0.69	0.70	0.78	0.48	0.80	0.81
[3]	0.58	0.92	0.86	0.84	0.86	0.84	78	Community	Twice a week	P	0.77	0.83	0.91	0.58	0.88	0.88
[4]	0.74	0.90	0.80	0.78	0.93	0.93	71	Students	7d	P	0.60	0.71	0.65	0.50	0.70	0.70
[5]	0.75	0.88	0.84	0.72	0.94	0.93	—	—	—	—	—	—	—	—	—	—
[6]	0.77	0.89	0.85	0.78	0.94	0.94	—	—	—	—	—	—	—	—	—	—
[7]	0.71	0.89	0.79	0.80	0.93	0.92	55	Students	7d	S	0.74	0.82	0.63	0.43	0.83	0.82
[8]	0.79	0.87	0.87	0.66	0.94	0.94	204	Students	2w	S	0.75	0.77	0.75	0.48	0.80	0.71
[9]	0.71	0.87	0.77	0.77	0.92	0.91	—	—	—	—	—	—	—	—	—	—
[10]	0.68	0.89	0.81	0.83	0.94	—	—	Patients	4w	—	0.67	0.82	0.72	0.66	0.85	—
[11]	0.81	0.86	0.82	0.67	0.93	—	57	Students	7d	S	0.68	0.72	0.77	0.60	0.78	0.79
[12]	0.70	0.87	0.84	0.81	0.94	0.93	81	Students	2w	S	0.63	0.69	0.70	0.35	0.76	0.76
[13]	0.63	0.88	0.76	0.72	0.92	0.91	14	Patients	7d	S	0.90	0.57	0.81	0.18	0.76	—
[14]	0.67	0.88	0.77	0.85	0.93	0.92	67	Students, community	2w	S	0.70	0.66	0.70	0.56	0.75	0.75
[15]	0.81	0.90	0.85	0.77	0.94	0.94	78	Students, community	2w+	S	0.76	0.85	0.79	0.45	0.87	0.87
[16]	0.76	0.88	0.85	0.76	0.94	0.93	70	Students	2w+	S	0.80	0.80	0.81	0.64	0.85	0.86
[17]	0.64	0.86	0.80	0.71	0.93	—	—	—	—	—	—	—	—	—	—	—
[18]	0.76	0.89	0.84	0.73	0.94	.94	—	—	—	—	—	—	—	—	—	—

Note. In the text, we refer to the numbers in the first column. CORE-OM: Clinical Outcome in Routine Evaluation–Outcome Measure; No.: number for indication in text; W: well-being; P: problems; F: functioning; R: risk; T: total score; -R: total score for nonrisk items; *N*: sample size; Type: sample type; Timespan: retest period in days (d) and weeks (w); Coef: coefficient (S: Spearman's ρ, P: Pearson's r); —: information not available.

our analyses. Table 6 summarizes the results of the psychometric analyses (see section Internal consistency [Cronbach's α]). The mean coefficients per scale (column 3) ranged from 0.93 (*total*) to 0.72 (*well-being*). The *well-being* and the *risk* scales showed lower values compared with the other scales and the total of items. All but the *problems* scale have significant heterogeneity values and a high amount (I^2 > 75%) of between studies' variance. The results of the outlier tests revealed two studies contributing most to variability in results [3, 5]. Moreover, the Croatian translation [3] contributed significantly to the mean Cronbach's α due to lower values in total score and total score of nonrisk items. Nevertheless, the values of internal consistency of the Croatian CORE-OM were still acceptable with $\overline{\alpha}_{total}$ = 0.86, confidence interval (CI) [0.82, 0.89] and $\overline{\alpha}_{total-R}$ = 0.84, CI [0.79, 0.88]. The Dutch translation [5] influenced significantly the mean Cronbach's α of the scale *problems* with $\overline{\alpha}_P$ = 0.88, CI [0.88, 0.88].

3.4 | Test stability

The shortest retest period was twice within 1 week [3]. The other retest periods were 1 week [2, 4, 7, 11, 13], 2 weeks or more [15, 16], and 1 month [10]. Six studies lacked information on retest periods, five papers provided no test stability analyses; seven studies used a student sample for retesting, two studies used community samples, and two used clinical samples for assessing test stability. We found two papers with a mixed sample of students and community members. The pooled test stability coefficients (see Table 6, section Test–retest reliability [Spearman's ρ]) ranged from $\overline{\rho}$ = 0.51 (*risk*) to $\overline{\rho}$ = 0.82 (*total*). The *risk* scale showed generally a low test stability and low heterogeneity between the studies; *well-being*, *problems*, *functioning*, and the total score as well as the total score of nonrisk items ranged from I^2 = 48% to I^2 = 72%. The Croatian CORE-OM [3] influenced significantly the pooled results of the *functioning* scale.

3.5 | Factorial structure

Table 4 shows results regarding the factorial structure of the CORE-OM translations. None of the studies applying factor analysis could replicate the intended four-factor structure of the instrument. Eight studies [1, 2, 4, 6, 7, 9, 11, 12] applied a PCA like Evans et al. (2002), four of which favoured three major components. The results were largely declared comparable with those of Evans et al., finding in almost all analyses a positively formulated domain measuring strengths, a negatively formulated domain measuring weaknesses, and the set of risk items. One study [2] suggested either a one-factorial or a two-factorial solution to describe their data adequately.

The four studies [3, 10, 12, 13] applying a CFA to assess the adequacy of the four-factorial structure of Evans et al. (2002) reported generally rather moderate results (Table 4). None of the studies applying a CFA found the four-factorial structure to describe the data best.

TABLE 4 Factorial structure of the CORE-OM translations

	Factorial structure									
	Principal component analysis				Confirmatory analysis					
No.	Sample	Rotation	Component	Factor	Sample	χ^2	*df*	GFI	RMSEA [CI]	CFI
[1]	C, NC	Oblique	3	POS, NEG, RISK	—	—	—	—	—	—
[2]	C, NC	Orthogonal	Uncertain	One global scale solution	—	—	—	—	—	—
[3]	—	—	—	—	NC	1,641.1	509	0.80	0.07 [—, —]	0.80
[4]	C, NC	Oblimin	3	POS, NEG, RISK	—	—	—	—	—	—
[5]	—	—	—	—	—	—	—	—	—	—
[6]	NC	Oblimin	3	POS, NEG, RISK	—	—	—	—	—	—
[7]	C, NC	Oblique	3	POS, NEG, RISK	—	—	—	—	—	—
[8]	—	—	—	—	—	—	—	—	—	—
[9]	C	—	3	—	—	—	—	—	—	—
[10]	—	—	—	—	C	—	—	0.88	0.06	—
[11]	C, NC	Oblimin	3	—	—	—	—	—	—	—
[12]	C, NC	Promax	2	Psychological distress, risk	NC	1,854.7	521	—	0.08 [—, —]	0.94
[13]	—	—	—	—	C, NC	3,964.2	561	—	0.057 [0.05, 0.06]	0.81
[14]	—	—	—	—	—	—	—	—	—	—
[15]	—	—	—	—	—	—	—	—	—	—
[16]	—	—	—	—	—	—	—	—	—	—
[17]	—	—	—	—	—	—	—	—	—	—
[18]	—	—	—	—	—	—	—	—	—	—

Note. In the text, we refer to the numbers in the first column. CORE-OM: Clinical Outcome in Routine Evaluation–Outcome Measure; No.: number for indication in text; Sample: C: clinical; NC: nonclinical; rotation: method to perform the factor extraction; components: number of extracted components; factors: contextual meaning of extracted factors; POS: positively worded items; NEG: negatively worded items; RISK: risk items; χ^2: Chi-square value; *df*: degrees of freedom; GFI: goodness of fit index; RMSEA: root mean square error of approximation; CI: confidence interval; CFI: comparative fit index; —: information not available.

3.6 | Convergent validity

Evans et al. (2002) correlated the English CORE-OM with well-established instruments. However, these instruments are not available in all target languages; thus, only nine studies deal with convergent validity. In total, six studies compared the CORE-OM with the SCL-90 and five with the BDI-II. Table 5 summarizes coefficients resulted from single studies, and Table 6 (lower part) shows the results of the validity analyses. The pooled correlation coefficients of the SCL-90 total score and the four scales ranged from $\bar{r}$ = 0.61 (*risk*) to $\bar{r}$ =0.82 (*total*).

All but the *problems* scale showed nonsignificant heterogeneity tests for both SCL-90 and BDI-II. Nevertheless, some studies showed a major influence on the variability between the studies (see Table 6, section Convergent validity [SCL-90]). The German CORE-OM [7] had a significant effect on the score for nonrisk items; the Albanian CORE-OM [2] contributed significantly to differences in variability between the studies in the scales *well-being* and *functioning*; and the results of the English CORE-OM [1] influence significantly the variability of the validity coefficients in the scale *risk* with $\bar{r}$ = 0.83, CI [0.68, 0.91].

The correlation coefficients of the CORE-OM scales with the total score of the BDI-II (see Table 6, section Convergent validity [BDI-II]) ranged from $\bar{r}$ = 0.53 (*risk*) to $\bar{r}$ = 0.84 (*total*). The heterogeneity coefficients varied from I^2 = 0% (*total*, *nonrisk items*, and *problems*) to I^2 = 52.9% (*well-being*). None of the results were significant. The Icelandic CORE-OM [8] contributed to the findings in the *functioning* scale. The Albanian CORE-OM [2] influenced the pooled correlation coefficients of the *well-being* and *functioning* scales.

3.7 | Moderator analysis

We found a moderating effect of the sample type upon Cronbach's α: The inpatient studies showed significantly lower internal consistency than the outpatient samples with respect to the total score, the nonrisk items, the *well-being* scale, and the *problems* scale (see Table 7). The coefficients in mixed samples did not differ significantly from the outpatient samples. None of the other moderator analyses revealed significant effects; therefore, they will not be reported in detail.

3.8 | Objectivity

The objectivity is difficult to evaluate in the face of the specificities of the target languages. All the versions of the CORE-OM have the same appearance (a two-sided A4 paper) with some slight optical differences (font type and size must be chosen from a list of given fonts). The head of the questionnaire containing sociodemographic data, treatment setting (beginning/follow-up/end of the treatment), and the instructions is a part of the translation process and should also be discussed in the focus group. On the bottom of the back page, there is space to provide calculation of the scale means and the sum of the total score as well as the total score of nonrisk items. The instructions at the top support comparability across test coordinators.

TABLE 5 Convergent validity of the CORE-OM translations

Convergent validity																	
	BDI-II								SCL-90								
No.	A/U	*N*	W	P	F	R	T	-R	A/U	*N*	W	P	F	R	T	-R	
[1]	++	29	0.79	0.74	0.78	0.32	0.81	0.83	++	34	0.68	0.87	0.79	0.83	0.88	0.85	
[2]	++	209	0.68	0.76	0.71	0.57	0.84	0.84	++	209	0.57	0.77	0.64	0.66	0.82	0.79	
[3]	++	—	—	—	—	—	0.79	—	+−	—	—	—	—	—	—	—	
[4]	−	—	—	—	—	—	—	—	+−	—	0.71	0.81	0.71	0.56	0.84	0.83	
[5]	+−	—	—	—	—	—	—	—	+−	—	—	—	—	—	—	—	
[6]	+−	—	—	—	—	—	—	—	++	201	0.67	0.83	0.73	0.57	0.81	0.82	
[7]	+−	—	—	—	—	—	—	—	++	135	0.71	0.85	0.77	0.58	0.86	0.86	
[8]	++	577	0.78	0.78	0.79	0.52	0.85	0.82	−	—	—	—	—	—	—	—	
[9]	−	—	—	—	—	—	—	—	+−	—	—	—	—	—	—	—	
[10]	+−	—	—	—	—	—	—	—	+−	—	—	—	—	—	—	—	
[11]	+−	—	—	—	—	—	—	—	−	—	—	—	—	—	—	—	
[12]	+−	—	—	—	—	—	—	—	+−	—	—	—	—	—	—	—	
[13]	+−	—	—	—	—	—	—	—	+−	—	—	—	—	—	—	—	
[14]	++	40	0.79	0.77	0.76	0.64	0.87	0.84	++	40	0.79	0.76	0.73	0.58	0.83	0.84	
[15]	++	162	0.79	0.80	0.74	0.48	0.83	0.83	++	155	0.70	0.77	0.72	0.46	0.79	0.79	
[16]	+−	—	—	—	—	—	—	—	+−	—	—	—	—	—	—	—	
[17]	+−	—	—	—	—	—	—	—	−	—	—	—	—	—	—	—	
[18]	+−	—	—	—	—	—	—	—	−	—	—	—	—	—	—	—	

Note. In the text, we refer to the numbers in the first column. CORE-OM: Clinical Outcome in Routine Evaluation–Outcome Measure; A/U: indicates whether a translation of the Beck Depression Inventory-II (BDI-II) or the Symptom Checklist-90 (SCL-90) is available (A) or not (U) in the respective language and whether it has been used; ++: exists and used; +−: exists and not used; −−: does not exist or has been established after publication of respective study. No.: number for indication in text; *N*: sample size; W: well-being; P: problems; F: functioning; R: risk; T: total score; -R: total score for nonrisk items; —: information not available.

TABLE 6 Results of heterogeneity tests

Scale	No. of items	α	ρ	r	CI [lower, upper]	Q	df	p	I^2 (%)	H^2	Papers contributing to differences
Internal consistency (Cronbach's α)											
Total score	34	0.93			[0.93, 0.94]	56.73	17	<0.001	81.26	5.34	[3]
Nonrisk items	28	0.92			[0.92, 0.94]	64.84	14	<0.001	88.24	8.50	[3]
Well-being	4	0.72			[0.69, 0.75]	86.20	17	<0.001	87.33	7.90	—
Problems	12	0.88			[0.88, 0.89]	19.08	17	0.32	6.25	1.07	[5]
Functioning	12	0.83			[0.81, 0.84]	76.83	17	<0.001	83.44	6.04	—
Risk	6	0.77			[0.74, 0.80]	186.64	17	<0.001	88.94	9.04	—
Test–retest reliability (Spearman's ρ)											
Total score	34		0.82		[0.78, 0.85]	22.07	11	0.02	52.44	2.10	—
Nonrisk items	28		0.81		[0.77, 0.85]	34.53	10	<0.001	69.93	3.33	—
Well-being	4		0.74		[0.69, 0.78]	22.42	11	0.02	48.56	1.94	—
Problems	12		0.77		[0.72, 0.81]	22.93	11	0.02	52.94	2.12	—
Functioning	12		0.78		[0.72, 0.82]	36.25	11	<0.001	72.01	0.57	[3]
Risk	6		0.51		[0.46, 0.55]	12.18	11	0.35	0.01	1.00	—
Convergent validity (SCL-90)											
Total score	34			0.82	[0.80, 0.84]	5.20	5	0.39	0.81	1.01	[7]
Nonrisk items	28			0.81	[0.80, 0.84]	5.50	5	0.36	19.88	1.25	[7]
Well-being	4			0.68	[0.62, 0.73]	9.24	5	0.1	46.48	1.87	[2]
Problems	12			0.81	[0.77, 0.83]	8.89	5	0.11	44.10	1.79	—
Functioning	12			0.72	[0.67, 0.76]	7.17	5	0.21	36.58	1.58	[2]
Risk	6			0.61	[0.51, 0.69]	15.96	5	0.01	72.61	3.65	[1]
Convergent validity (BDI-II)											
Total score	34			0.84	[0.82, 0.86]	1.21	4	0.88	0.00	1.00	—
Nonrisk items	28			0.83	[0.81, 0.85]	0.61	4	0.96	0.00	1.00	—
Well-being	4			0.76	[0.71, 0.81]	7.90	4	0.1	52.90	2.12	[2]
Problems	12			0.78	[0.75, 0.80]	1.17	4	0.88	0.00	1.00	—
Functioning	12			0.76	[0.71, 0.79]	4.94	4	0.29	34.67	1.53	[2, 8]
Risk	6			0.53	[0.47, 0.57]	4.22	4	0.38	0.02	1.00	—

Note. BDI-II: Beck Depression Inventory-II; SCL-90: Symptom Checklist-90; CI: confidence interval [lower, upper]; *Q*: Cochran's *Q*; *df*: degrees of freedom; *p*: *p* value; I^2 and H^2: heterogeneity coefficients; —: none.

TABLE 7 Results of the moderator analysis for Cronbach's α by sample type

Cronbach's α										
	Omnibus test				Inpatients[a]			Inpatients and outpatients (mixed sample)[a]		
Scale	*k*	Q_M	*df*	*p*	β	*z*	*p*	β	*z*	*p*
Total score	18	4.30	2	0.12	−0.12	−2.02	0.04	−0.06	−0.89	0.38
Nonrisk items	15	7.45	2	0.02	−0.21	−2.72	0.01	−0.06	−0.80	0.42
Well-being	18	5.51	2	0.06	−0.16	−2.34	0.02	−0.30	−0.38	0.71
Problems	18	7.17	2	0.03	0.06	2.55	0.01	−0.02	−0.61	0.54
Functioning	18	3.94	2	0.14	−0.09	−1.50	0.13	−0.12	−1.60	0.11
Risk	18	1.57	2	0.46	0.09	1.18	0.24	0.07	0.69	0.49

Note. *k*: number of studies; Q_M: empirical *Q* value for moderator analysis; *df*: degrees of freedom; *p*: *p* value; β: regression coefficient; *z*: *z* value.

[a]Compared with the outpatient sample.

The instructions do not provide specific details regarding the interpretation of the results. The manual of the English CORE-OM contains cut-off points indicating the CSC. To allow for comparing different versions of the CORE-OM, we need standardized scores, norm tables, and reference values for CSC (split by gender, age, or other relevant factors), which did not appear in the reviewed studies.

4 | DISCUSSION

4.1 | Summary

The present study compared systematically 21 translations of the CORE-OM applying methods of systematic review and meta-analysis.

The research focussed on the comparability of the translations with respect to psychometric properties (especially reliability, validity, and objectivity). Our results show that the different versions of the CORE-OM are largely comparable from a psychometric point of view and adequately reflect the English CORE-OM. Despite a certain heterogeneity in data collection, sample sizes, and composition of the samples, the international versions of the CORE-OM provide similar results of psychometric analyses regarding all criteria.

We identified six studies contributing significantly to internal consistency, retest reliability, and convergent validity, four of which in a positive direction and two (Croatia and Albania) showed significantly lower values in internal consistency and convergent validity, respectively. However, the internal consistency of the Croatian version was still above 0.80 and therefore satisfactory. In contrast, the Albanian *well-being* and *functioning* scales had significantly and severely lower correlations with both external criteria (SCL-90 and BDI-II). This may be due to poor translation of the CORE-OM, poor translation of the external criteria, inappropriate samples, or a specific attitude towards the well-being and functioning constructs in the Albanian context. This should be pursued further in targeted studies.

The CORE-OM was developed for outpatients (Barkham et al., 1998). Because some studies collected inpatient data as well, we could compare the sample types in a moderator analysis. The internal consistency of inpatient samples was significantly lower than that of the outpatient samples; that is, the CORE-OM performs better in the sample type it has been developed for. Hence, it should be used cautiously in inpatient settings, particularly in multicentre studies.

4.2 | Reported analyses

Almost all of the 21 studies (18) analysed internal consistency, and 13 assessed retest reliability. Only nine translations performed a validation using external measures such as BDI-II and SCL-90, although these two instruments are available in almost all target languages (see Table 4). Likewise, the examination of the factorial validity was seldom carried out (7 PCA; 4 CFA). Therefore, factorial validity is difficult to evaluate. This may be due to the fact that the current translation guidelines (CST, 2015) do not provide any details regarding the assessment of validity, which we would consider a worthwhile extension.

4.3 | Sampling

Most of the samples were convenience samples for both clinical and nonclinical populations. Additionally, half of the studies assessing stability used student samples. Therefore, a generalization of these results is only possible with great caution, if at all.

4.4 | Factorial structure

None of the studies—including Evans et al. (2002)—could replicate the originally proposed four-factorial structure of the CORE-OM. Rather, the majority of the studies suggested that the 34 items represent a three-factorial latent structure: positively worded items (assessing strengths), negatively worded items (assessing disabilities and distress), and the items of the *risk* scale. These findings are in line with former results regarding the factorial structure of the English CORE-OM: Even the PCA of Evans et al. suggested a three-factor solution. Lyne, Barrett, Evans, and Barkham (2006) suggested a two-factorial structure (risk and psychological distress), recommending to use the risk items as a separate indicator of risky and self-harming behaviour, but only by professionals. Handscomb, Hall, Hoare, and Shorter (2016) applied a CFA in a sample of tinnitus patients. They estimated 10 different model variants derived from previous studies on the CORE-OM, finding also the poorest fit for the original four-factorial solution and the best fit for the model containing negative, positive, and risk factors (i.e., the one that had already been identified by Evans et al., 2002). Nevertheless, the questionnaire remained unaltered with respect to both number of items and scoring. Because the translated versions of the instrument have adopted this deficiency, we see a clear need for further research on the factorial structure and scoring of the CORE-OM.

4.5 | The *risk* scale

The results indicate severe problems of the *risk* scale. The risk construct itself seems to function poorly in the selected clinical population across all countries. We have to assume that patients treated in an outpatient setting (for which the CORE-OM has been designed) have already reached a certain degree of stability and are, therefore, not acutely at risk. None of the studies analysed here reported information on medical care for patients.

Hence, this scale seems applicable rather in an (inpatient) psychiatric setting, for example, shortly after admission and at the end of the stay, to detect possible changes. The psychometric properties of the *risk* scale in studies using inpatient samples (i.e., Croatian, Czech, Russian, and Slovak CORE-OM) did not differ significantly from the studies using outpatients. Therefore, the *risk* scale requires further detailed research. Evans et al. (2002) have already suggested evaluating the total score without considering risk items, and the present research indicates again that this calculation approach should be pursued further. For example, studies involving patients with potentially harmful behaviour (e.g., drugs or substance abuse, psychoses, and depression) could clarify the mediocre results of this scale.

4.6 | Independence with respect to diagnostic groups

Evans et al. (2000) considered the CORE-OM suitable for all diagnostic groups (p. 253), which seems questionable in the light of our empirical results. There are some publications on special disorder groups investigating the psychometric properties of the CORE-OM, such as eating disorder (Jenkins & Turner, 2014) or emotional distress in people with tinnitus (Handscomb et al., 2016). But we could not identify studies dedicated to the applicability of the CORE-OM to patients with personality disorder or psychoses.

4.7 | Reporting standards

Our review showed also the need for more specific guidelines for reporting the results of psychometric analyses. The current CST (2015) guidelines specify detailed steps regarding the translation processes, so that translating authors are subject to highly standardised procedures. In contrast, no specific guidelines exist regarding mandatory analyses and reporting standards. It should be determined which methods are suitable for recording the respective psychometric properties (e.g., whether Spearman or Pearson correlation coefficients should be preferred for assessing the test stability). The sample sizes should be supported by a power analysis to clarify the consequences of noncompliance with regulations. The sampling requires clear presentation (see Lounsbury, Gibson, & Saudargas, 2006, discussing the consequences of using student samples). Furthermore, a standardized presentation of the results would increase the comparability of studies. Our results show further that the studies dealt only marginally with the calculation of both the CSC and the reliable change. Because the CORE-OM is primarily suitable for measuring change, the necessity of both indices seems highly indicated.

4.8 | Methodology

We consider the chosen procedure appropriate for comparing the various translations. Gilbody, Richards, Brealey, and Hewitt (2007) conducted a similar analysis using international versions of the Patient Health Questionnaire (Spitzer, Kroenke, & Williams, 1999) and Patient Health Questionnaire-9 (Kroenke, Spitzer, & Williams, 2001). This technique has proven successful for comparative studies on questionnaires and can, therefore, be considered as a standard procedure for multiple language translation.

4.9 | Conclusion

The question, whether different translations of the CORE-OM can be treated as one and the same instrument, could therefore be answered with "yes." However, reservations exist regarding the quality of the original (English) CORE-OM, especially regarding the factorial structure. All translations applied the original factorial structure thus adopting its weaknesses as well. Therefore, we recommend a revision of the instrument in this regard. Keeping in mind that we dispose already of numerous follow-up studies probing various alternative models, it is interesting to note that none of these results has so far found its way into the CORE-OM. A very promising candidate was the approach of Lyne et al. (2006). The authors used a "nested factors first-order general factor model with four residualized latents (…) and with two method latents of positively and negatively worded items" (p. 195). However, this complex model would not allow for a straightforward scoring required in a clinical daily routine. Another promising candidate would be the three-factorial model of Evans et al. (2002), which deserved a closer inspection, possibly involving item response theory models (e.g., de Ayala, 2009).

Our results show further that the instrument performs better with outpatient samples, which has to be considered when using the CORE-OM in multicentre studies. Finally, international guidelines for the reporting on translation and adaptation studies should be established. This will increase both the quality of the studies and the comparability between different translations.

DECLARATION OF INTEREST STATEMENT

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The authors thank Alexandra Maria Kratki and Linus Christoph Winter for their help in the literature research.

ORCID

Marina Zeldovich https://orcid.org/0000-0003-0172-9904
Rainer W. Alexandrowicz https://orcid.org/0000-0001-9846-8936

REFERENCES

Andrews, G., Peters, L., & Teesson, M. (1994). *The measurement of consumer outcome in mental health: A report to the National Mental Health Information Strategy Committee*. Sydney: Clinical Research Unit for Anxiety Disorders.

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*(1), 35–47. https://doi.org/10.1080/09638239818328

Barkham, M., Gillian, E. H., & Startup, M. (1996). The IIP-32: A short version of the Inventory of Interpersonal Problems. *Clinical Psychology*, *35*(1), 21–35. https://doi.org10.1111/j.2044-8260.1996.tb01159.x

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*(1), 3–15. https://doi.org/10.1080/14733140600581218

Baujat, B., Mah, C. D., Pignon, J.-P., & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine*, *21*, 2641–2652. https://doi.org/10.1002/sim.1221

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II (BDI-II)*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Bodinaku, B. (2014). Translation, validation and standardization of the Albanian version of the SCL-90-R (Symptom Checklist-90-Revised) and CORE-OM (Clinical Outcomes in Routine Evaluations – Outcome Measure). Sigmund Freud Private University, Vienna. (Doctoral thesis)

Brook, R. H., Williams, K. N., & Avery, A. D. (1976). Quality assurance today and tomorrow: Forecast for the future. *Annals of Internal Medicine*, *85*(6), 809–817. https://doi.org/10.7326/0003-4819-85-6-809

Campbell, M. M. (2011). Introducing the CORE-OM in a South African context: Validation of the CORE-OM using a South African student population sample. *South Africa Journal of Psychology*, *41*(4), 488–502. https://doi.org/10.1177/008124631104100408

Campbell, M. M., & Young, C. (2016). A Xhosa language translation of the CORE-OM using South African university student samples. *Transcultural Psychiatry*, *53*(5), 64–74. https://doi.org/10.1177/1363461516661643

Cartasso, G. & Lemos, A. (2012). Una investigación a largo plazo en un espacio de escucha facilitado desde el enfoque centrado en la persona. [A long-term research into a listening space provided from the person centered approach]. In *Viii congreso de counseling de las américas*.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, *37*(3), 256–266. https://doi.org/10.2307/2332378

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Core System Trust (2011, April 1). Translating and "normalising" CORE system CORE System Trust (CST) position statement. Retrieved from https://www.psyctc.org/CORE-OM/Translating_CORE-OM_CST_position_statement.pdf.

Core System Trust (2015, April 1). Translation policy. Retrieved from https://www.coresystemtrust.org.uk/cst-translation-policy/.

Core System Trust (2018, January 24). Retrieved from www.coresystemtrust.org.uk/translations.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. NY: Guilford.

Derogatis, L. R. (1983). *SCL-90-R: Administration, Scoring, & Procedures: Manual*. Towson, MD: Clinical Psychometric Research.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13*, 595–605. https://doi.org/10.1017/S0033291700048017

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

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*(5), 447–455. https://doi.org/10.1002/cpp.1788

Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The Global Assessment Scale: A procedure of measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry*, *33*, 766–771. https://doi.org/10.1001/archpsyc.1976.01770060086012

Erickson, R. C. (1975). Outcome studies in mental hospitals: A review. *Psychological Bulletin*, *82*(4), 519–540. https://doi.org/10.1037/h0076899

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

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence-Based Mental Health*. Royal College of Psychiatrists, *1*(3), 70–72. https://doi.org/10.1136/ebmh.1.3.70

Evans, C., Mellor-Clark, J., Marginson, F., & Barkham, M. (2000). CORE: Clinical Outcomes in Routine Evaluation. *Journal of Metal Health*, *9*(3), 247–255. https://doi.org/10.1080/jmh.9.3.247.255

Feixas, G., Evans, C., Trujillo, A., Saul, L. A., Botella, L., Corbella, S., … López-González, M. (2012). La versión española del CORE-OM: Clinical Outcomes in Routine Evaluation Outcome Measure. *Aportaciones Teoricas e Instrumentales*, *23*(89), 109–135.

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). robumeta: Robust variance meta-regression. R package version 2.0. Retrieved from https://CRAN.R-project.org/package=robumeta.

Flanagan, J. C. (1978). A research approach to improving our quality of life. *American Psychologist*, *33*, 138–147. https://doi.org/10.1037//0003-066X.33.2.138

Gampe, K., Bieščad, M., Balúnová-Labanicová, L., Timulák, L., & Evans, C. (2012). Slovenská adaptácia metódy CORE-OM [Slovak adaptation of the CORE-OM]. *Ceská a slovenská psychiatrie.[Czech and Slovak Psychiatry]*, *39*(1), 4–13.

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine*, *22*(11), 1596–1602. https://doi.org/10.1007/s11606-007-0333-y

Goldberg, D. P., & Hiller, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine*, *9*, 139–145. https://doi.org/10.1017/S0033291700021644

Goldhahn, J., Shisha, T., Macdermid, J. C., & Goldhahn, S. (2013). Multilingual cross-cultural adaptation of the patient-rated wrist evaluation (PRWE) into Czech, French, Hungarian, Italian, Portuguese (Brazil), Russian and Ukrainian. *Archives of Orthopaedic and Trauma Surgery*, *133*(5), 589–593. https://doi.org/10.1007/s00402-013-1694-9

Handscomb, L., Hall, D. A., Hoare, D. J., & Shorter, G. W. (2016). Confirmatory factor analysis of Clinical Outcomes in Routine Evaluation (CORE-OM) used as a measure of emotional distress in people with tinnitus. *Health and Quality of Life Outcomes*, *14*(124), 1–9. https://doi.org/10.1186/s12955-016-0524-5

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558. https://doi.org/10.1002/sim.1186

Honkalampi, K., Laitila, A., Juntunen, H., Lehmus, K., Piiparinen, A., Törmänen, M. K., & Evans, C. (2017). The Finnish Clinical Outcome in Routine Evaluation Outcome Measure: Psychometric exploration in clinical and non-clinical samples. *Nordic Journal of Psychiatry*, *71*(8), 589–597. https://doi.org/10.1080/08039488.2017.1365378

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). Inventory of Interpersonal Problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, *56*, 885–892. https://doi.org/10.1037/0022-006X.56.6.885

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336–352. https://doi.org/10.1016/S0005-7894(84)80002-7

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

Jenkins, P. E., & Turner, H. M. (2014). An investigation into the psychometric properties of the COREOM in patients with eating disorders. *Counselling and Psychotherapy Research*, *14*(2), 102–110. https://doi.org/10.1080/14733145.2013.782057

Jokic-Begic, N., Korajlija, A. L., & Jurin, T. (2014). Faktorska struktura, psihometrijske karakteristike i kritična vrijednost hrvatskoga prijevoda CORE-OM upitnika [Factorial structure, psychometric properties and critical values of the Croatian version of the CORE-OM]. *Psihologijske Teme*, *23*(2), 265–288.

Juhová, D. (2015). Adaptace metod CORE-OM a ORS do českého prostředí [Adaptation of the CORE-OM and the ORS to Czech]. Masaryk University Brno. (Diploma thesis)

Juntunen, H., Piiparinen, A., Honkalampi, K., Inkinen, M., & Laitila, A. (2015). CORE-OM-mittarin suomalainen validointitutkimus yleisväestössä. [The Finnish Validation Study of the CORE-OM-measure: Non-clinical sample]. *Psykologia*, *50*(4), 109–135.

Kaplan, R. M., & Bush, J. W. (1982). Health-related quality of life measurement for evaluation research and policy analysis. *Health Psychology*, *1*, 61–80. https://doi.org/10.1037/0278-6133.1.1.61

Kristjánsdóttir, H., Sigurðsson, B. H., Salkovskis, P., Ólason, D., Sigurdsson, E., Evans, C., … Sigurðsson, J. F. (2015). Evaluation of the psychometric properties of the Icelandic version of the Clinical Outcomes in Routine Evaluation – Outcome Measure, its transdiagnostic utility and cross-

cultural validation. *Clinical Psychology and Psychotherapy*, *22*, 64–74. https://doi.org/10.1002/cpp.1874

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*, 606–613. https://doi.org/10.1046/j.15251497.2001.016009606.x

Lewis, S., & Clarke, M. (2001). Forest plots: Trying to see the wood and the trees. *BMJ: British Medical Journal*, *322*(7300), 1479–1480. https://doi.org/10.1136/bmj.322.7300.1479

Lohr, N. K. (1988). Outcome measurement: Concepts and questions. *Inquiry*, *25*(1), 37–50.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Lounsbury, J. W., Gibson, L. W., & Saudargas, R. A. (2006). Scale development. In F. T. L. Leong, & J. T. Austin (Eds.), *The psychology research handbook: A guide for graduate students and research assistants* (2nd ed.) (pp. 125–136). Thousand Oaks: Sage Publications. https://doi.org/10.4135/9781412976626.n9

Lyne, K. J., Barrett, P., Evans, C., & Barkham, M. (2006). Dimensions of variation on the CORE-OM. *British Journal of Clinical Psychological Society*, *45*, 185–203. https://doi.org/10.1136/bmj.322.7300.1479

Meerding, W. J., van't Spijker, A., & van Riessen, M. (2012). Monitoren van behandelresultaat met de CORE-OM. Bruikbaarheid en psychometrische eigenschappen. [Monitoring of treatment results using the CORE-OM. Usability and psychometric properties.]. *Tijdschrift voor Psychotherapie*, *38*(5), 355–367. https://doi.org/10.1007/s12485-012-0041-x

Mellor-Clark, J., Barkham, M., Connell, J., & Evans, C. (1999). Practice-based evidence and standardized evaluation: Informing the design of the CORE system. *The European Journal of Psychotherapy, Counselling & Health*, *2*(3), 357–374. https://doi.org/10.1080/13642539908400818

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … PRISMA-P Group (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*, 1. https://doi.org/10.1186/2046-4053-4-1

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. M. (2002). The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *The British Journal of Psychiatry*, *180*(5), 461–464. https://doi.org/10.1192/bjp.180.5.461

Palmieri, G., Evans, C., Hansen, V., Brancaleon, 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. https://doi.org/10.1002/cpp.646

Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology*, *6*(1549), 1–9. https://doi.org/10.3389/fpsyg.2015.01549

R Core Team (2013). R: A language and environment for statistical computing [Software manual]. Vienna, Austria. Retrieved from www.r-project.org.

Sales, C. M. D., de Matos Moleiro, C. M., Evans, C., & Alves, P. C. G. (2012). Versão Portuguesa do CORE-OM: tradução, adaptação e estudo preliminar das suas propriedades psicométricas. [The Portuguese version of the CORE-OM: Translation, adaptation, and preliminary study of psychometric properties]. *Revista de Psiquiatria Clínica*, *39*(2), 54–59. https://doi.org/10.1590/S0101-60832012000200003

Santana, M. R. M., da Silva, M. M., de Moraes, D. S., Fukunda, C. C., Freitas, L. H., Ramos, M. E. C., et al. (2015). Brazilian Portuguese version of the CORE-OM: Cross-cultural adaptation of an instrument to assess the efficacy and effectiveness of psychotherapy. *Trends Psychiatry Psychotherapy*, *37*(4), 227–231. https://doi.org/10.1590/2237-6089-2015-0002

Skre, I., Friborg, O., Elgaroy, S., Evans, C., Myklebust, L. H., Lillevoll, K., … Hansen, V. (2013). The factor structure and psychometric properties of the Clinical Outcome in Routine Evaluation–Outcome Measure (CORE-OM) in Norwegian clinical and non-clinical samples. *Biomed Central Psychiatry*, *13*, 13–99. https://doi.org/10.1186/1471-244X-13-99.

Slade, M. (1996). Assessing the needs of the severely mentally ill: Cultural and professional differences. *Australian and New Zealand Journal of Psychiatry*, *10*, 743–753. https://doi.org/10.1177/002076409604200101

Slade, M. (2002). What outcomes to measure in routine mental health services, and how to assess them: A systematic review. *International Journal of Social Psychiatry*, *100*, 149–157. https://doi.org/10.1046/j.1440-1614.2002.01099.x

Slade, M. (2010). Outcome measurement in England. In T. Trauer (Ed.), *Outcome measurement in mental health: Theory and praxis* (pp. 32–39). New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511760686.005

Slade, M., Leese, M., Taylor, R., & Thornicroft, G. (1999). The association between needs and quality of life in an epidemiologically representative sample of people in psychosis. *Acta Psychiatrica Scandinavica*, *100*, 149–157. https://doi.org/10.1111/j.1600-0447.1999.tb10836.x

Slade, M., Thornicroft, G., & Glover, G. (1999). The feasibility of routine outcome measures in mental health. *Social Psychiatry and Psychiatric Epidemiology*, *34*(5), 243–249. https://doi.org/10.1007/s001270050139

Spitzer, R. L., Kroenke, K., & Williams, J. B. W. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. Primary care evaluation of mental disorders. Patient Health Questionnaire. *JAMA*, *282*, 1734–1744. https://doi.org/10.1001/jama.282.18.1737

Sproll, S. (2011). Ein kurzes Outcome-Maß zur routinemäßigen Datenerhebung im Kontext von Evidence-Based Practice und Practice-Based Evidence in der Psychotherapie – Übersetzung und psychometrische Eigenschaften des deutschen CORE-OM. Sigmund Freud Private University, Vienna. (Diploma thesis)

Sutherland, H. J., & Till, J. E. (1993). Quality of life assessments and levels of decision making: differentiating objectives. *Quality of Life Research*, *2*, 297–303. https://doi.org/10.1007/BF00434801

Thornicroft, G., & Slade, M. (2014). New trends in assessing the outcomes of mental health interventions. *World Psychiatry*, *13*, 118–124. https://doi.org/10.1002/wps.20114

Thornicroft, G., & Tansella, M. (2010). In G. Thornicroft, & M. Tansella (Eds.), *Mental health outcome measures* (3rd ed.). London: RCPsych Publications.

Trujillo, A., Feixas, G., Bados, A., García-Grau, E., Salla, M., Medina, J. C., … Evans, C. (2016). Psychometric properties of the Spanish version of the Clinical Outcomes in Routine Evaluation–Outcome Measure. *Neuropsychiatric Disease and Treatment*, *12*, 1457–1466. https://doi.org/10.2147/NDT.S103079

Uji, M., Sakamoto, A., Adachi, K., & Kitamura, T. (2012). Psychometric properties of the Japanese version of the Clinical Outcomes in Routine Evaluation–Outcome Measure. *Comprehensive Psychiatry*, *53*, 600–608. https://doi.org/10.1016/j.comppsych.2011.09.006

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Viliuniene, R., Evans, C., Hilbig, J., Pakalniskiene, V., Danieleviciute, V., & Laurinaitis, E. (2012). Validation of the Lithuanian version of the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM). *Nordic Journal of Psychiatry*, *67*, 1–7. https://doi.org/10.3109/08039488.2012.745599

Wing, J. K., Beevor, A. S., Curtis, R. H., Park, S. G. B., Hadden, J., & Burns, A. (1998). Health of the Nation Outcome Scales (HoNOS): Research and development. *British Journal of Psychiatry*, *172*, 11–18. https://doi.org/10.1192/bjp.172.1.11

World Health Organization (2018). Process of translation and adaptation of instruments. Retrieved from http://www.who.int/substance_abuse/research_tools/translation/en/.

Zeldovich, M. A., Ivanov, A. A., Evans, C., & Andreas, S. (2014). Psyhometrichesky obzor russkoy versii oprosnika Clinical Outcome in Routine Evaluation – Outcome Measure (CORE-OM). [Psychometric properties of the Russian Version of the CORE-OM]. *Chelovechesky Kapital*, *1*(61), 58–67.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zeldovich M, Alexandrowicz RW. Comparing outcomes: The Clinical Outcome in Routine Evaluation from an international point of view. *Int J Methods Psychiatr Res*. 2019;28:e1774. https://doi.org/10.1002/mpr.1774