Psychological Assessment

Preliminary Validation of the Rating of Outcome Scale and Equivalence of Ultra-Brief Measures of Well-Being

Jason A. Seidel, William P. Andrews, Jesse Owen, Scott D. Miller, and Daniel L. Buccino

Online First Publication, April 21, 2016. http://dx.doi.org/10.1037/pas0000311

CITATION

Seidel, J. A., Andrews, W. P., Owen, J., Miller, S. D., & Buccino, D. L. (2016, April 21). Preliminary Validation of the Rating of Outcome Scale and Equivalence of Ultra-Brief Measures of Well-Being. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0000311

Psychological Assessment

© 2016 American Psychological Association
1040-3590/16/$12.00 http://dx.doi.org/10.1037/pas0000311

Preliminary Validation of the Rating of Outcome Scale and Equivalence of Ultra-Brief Measures of Well-Being

Jason A. Seidel
Colorado Center for Clinical Excellence, Denver, Colorado

William P. Andrews
Pragmatic Research Network, Manyother Ltd., Bangor, Wales

Jesse Owen
University of Denver

Scott D. Miller
International Center for Clinical Excellence, Chicago, Illinois

Daniel L. Buccino
Johns Hopkins Bayview Medical Center, Baltimore, Maryland

Three brief psychotherapy outcome measures were assessed for equivalence. The Rating of Outcome Scale (ROS), a 3-item patient-reported outcome measure, was evaluated for interitem consistency, test–retest reliability, discriminant validity, repeatability, sensitivity to change, and agreement with the Outcome Rating Scale (ORS) and Outcome Questionnaire (OQ) in 1 clinical sample and 3 community samples. Clinical cutoffs, reliable change indices, and Bland-Altman repeatability coefficients were calculated. Week-to-week change on each instrument was compared via repeated-measures-corrected effect size. Community-normed *T* scores and Bland-Altman plots were generated to aid comparisons between instruments. The ROS showed good psychometric properties, sensitivity to change in treatment, and discrimination between outpatients and nonpatients. Agreement between the ROS and ORS was good, but neither the agreement between these nor that between ultrabrief instruments and the OQ were as good as correlations might suggest. The ROS showed incremental advantages over the ORS: improvements in concordance with the OQ, better absolute reliability, and less oversensitivity to change. The ROS had high patient acceptance and usability, and scores showed good reliability, cross-instrument validity, and responsiveness to change for the routine monitoring of clinical outcomes.

Keywords: outcome measure, rating scale, reliable change, Bland-Altman, effect size

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Clinicians should integrate the "best available research evidence" with clinical expertise while "monitoring patient progress and adjusting practices accordingly" (APA Presidential Task Force on Evidence-Based Practice, 2006, pp. 273, 276; also see National Quality Forum, 2013). To aid clinicians in this endeavor, several brief patient-reported outcome measures (PROMs) are available at low or no cost for monitoring patient progress in psychotherapy. However, there is not yet widespread use of such measures (Harmon et al., 2007; Swift, Greenberg, Whipple, & Kominiak, 2012; Tarescavage & Ben-Porath, 2014).

Many clinicians have reported concerns about time burden and difficulty with interpreting scores from outcome measures (Tarescavage & Ben-Porath, 2014), with most clients and clinicians not tolerating even 5 min of outcomes assessment for session-by-session measurement (Australian Mental Health Outcomes and Classification Network, 2005; Miller, Duncan, Brown, Sparks, & Claud, 2003). Given the resource constraints faced in typical clinical settings, practitioners and clients would benefit from very brief, psychometrically sound, and more interpretable PROMs.

PROMs such as the Outcome Questionnaire—45.2 (OQ; Lambert et al., 2004), Outcome Rating Scale (ORS; Miller, Duncan, Brown, Sparks, & Claud, 2003; Miller & Duncan, 2004), Clinical Outcomes in Routine Evaluation (CORE) sys-

Jason A. Seidel, Colorado Center for Clinical Excellence, Denver, Colorado; William P. Andrews, Pragmatic Research Network, Manyother Ltd., Bangor, Wales; Jesse Owen, Counseling Psychology Department, University of Denver; Scott D. Miller, International Center for Clinical Excellence, Chicago, Illinois; Daniel L. Buccino, Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland.

Research was performed at the Colorado Center for Clinical Excellence. The authors wish to thank Bob Bertolino, PhD, for his invaluable help in the pilot phase of this project.

Jason Seidel is a developer of Rating of Outcome Scale. William Andrews is a developer of Pragmatic Tracker software platform for Rating of Outcome Scale, Outcome Rating Scale, CORE-10, and other outcome measures. Scott Miller developed and is the co-owner of the copyright for the Outcome and Session Rating Scales. Individual practitioners may obtain a free, lifetime license to use the scales by registering at http://scott-d-miller-ph-d.myshopify.com/collections/performance-metrics/products/performancemetrics-licenses-for-the-ors-and-srs. Researchers may obtain permission to use the ORS by writing to info@scottdmiller.com.

Correspondence concerning this article should be addressed to Jason Seidel, Colorado Center for Clinical Excellence, 1720 South Bellaire Street, Suite 204, Denver, CO 80222. E-mail: jseidel@thecoloradocenter.com

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

tem (including the CORE-OM and CORE-10; Connell & Barkham, 2007), and Behavioral Health Measure (Kopta & Lowry, 2002), were designed to assess changes in general psychological functioning over the course of therapy, and their psychometric properties have been supported in dozens of studies (e.g., A. Campbell & Hemsley, 2009; Connell & Barkham, 2007; Janse, Boezen-Hilberdink, van Dijk, Verbraak, & Hutschemaekers, 2014; Kopta & Lowry, 2002; Lambert et al., 2004; Reese, Norsworthy, & Rowlands, 2009). However, little attention has been paid to the agreement between measures. Most researchers in the field have relied on correlation coefficients to assess concurrent validity and test–retest reliability rather than score agreement and repeatability. Correlations are inappropriate because they measure the strength of association (relative reliability), not agreement or absolute reliability (Bland & Altman, 1986; Vaz, Falkmer, Passmore, Parsons, & Andreou, 2013). Correlations can be weak when scores agree strongly, and correlations can be strong while scores show poor agreement. What clinicians need to know is whether the scores from one instrument are equivalent to the scores from another instrument, not just whether (as with height and weight) they are strongly associated while being quite different. Bland-Altman plots (Bland & Altman, 1986) are used frequently in medical research and to a lesser extent in psychological research to show agreement between two methods for measuring a dependent variable. While some instrument comparisons (e.g., digital vs. mercury thermometers) are likely to have both precision (i.e., repeatable or similar, not merely correlated measurements) and high agreement with other instruments, comparisons between more complex measurement methods and variables (e.g., treatment response in leukemia; Müller et al., 2009) are likely to show limited repeatability and agreement. Along these lines, it is common for patient reports of subjective experiences such as pain to yield low test–retest repeatability and intermethod agreement (DeLoach, Higgins, Caplan, & Stiff, 1998; Lima, Barreto, & Assunção, 2012). Given the relative lack of precision in PROMs that measure complex subjective experiences, it is especially important to begin to quantify and—if possible—improve on the repeatability of scores and the agreement between ostensibly similar measures.

PROMs in routine practice vary in length (from 4 to over 60 items). The ORS is an ultrabrief measure (four items) developed especially with the intention of balancing brevity with validity, based on the 45-item OQ. While the ORS has been used widely, concerns include (a) the poor repeatability of visual analogue scales for subjective experiences, even when taken only 1 min apart, and greater interrater agreement with a Likert-type response format (Bijur, Silver, & Gallagher, 2001; DeLoach et al., 1998; Wuyts, De Bodt, & Van de Heyning, 1999); (b) no published analyses of agreement or repeatability for the ORS; and (c) no published analyses comparing the ORS's sensitivity to change with other instruments. Clearly, there is room for more than one ultrabrief measure for tracking client outcomes in therapy; but more importantly, further research is needed on the comparability of ultrabrief measures given their relatively higher rates of acceptance by clients and clinicians. Accordingly, the current study tested a new three-item measure called the Rating of Outcome Scale (ROS; see the Appendix; Seidel, 2011) in clinical and community samples, and compared it with the ORS and OQ.

Method

Participants

One clinical sample and three community samples were administered combinations of measures in counterbalanced order, either once or approximately 1 week apart. The clinical sample received all three measures at each session for up to three sessions. The Community1 sample received a single administration of the ROS and OQ; the Community2 sample received a single administration of all three measures, and the Community3 sample received three administrations of the ROS and ORS.

Clinical sample. The clinical sample was drawn from a private outpatient psychotherapy practice with multiple clinicians. Adult clients who came for at least one session of psychotherapy in their first treatment episode at the practice ($n = 279$) were informed that the practitioners routinely collected client data as an integral part of Feedback-Informed Treatment (FIT; see Bertolino & Miller, 2013). Eighty-six of the clients were administered the three measures in each session for up to three sessions. Of these 86, 78 clients attended two sessions and completed all measures, and 73 attended three sessions and completed all measures. The remaining 193 clients received the three measures only in the first session. Of the 279 participants, 89% identified as White or Caucasian, and 51.3% were female. The median education of the sample was 16 years (range = 8–26), approximately 26% of the clients in the sample sought reimbursement from their insurance or a health spending account, and approximately 11% paid a reduced fee or no fee. Per the general policy of the practice, *Diagnostic and Statistical Manual of Mental Disorders* and *International Statistical Classification of Diseases and Related Health Problems* diagnoses were rarely assigned in this general outpatient practice. Clients' own descriptions of their presenting problems (primarily from free-response questions on an intake questionnaire) were used to classify them for this study. Among the top five presenting problem areas, 62.1% of clients reported some kind of relational problem at intake; 27.2% reported stress, anxiety, or trauma concerns; 20.7% reported problems with mood or depression; 12.4% reported identity/role concerns or confidence issues; and 8.3% had anger problems. While 58.6% reported only one problem area at intake, 27.8% reported two problem areas, and 13.6% reported three or more major categories of distress at intake.

Three community samples. The first community sample (Community1; $n = 708$) was drawn from U.S.-based members of an online labor-market community (Amazon.com's Mechanical Turk, or Mturk) consisting of a pool of over 2 million workers who complete brief tasks in exchange for very small payments. The sample was not selected or screened for nonclinical levels of distress. Like the other community samples, they were asked about past and current psychotherapy or psychiatric medication treatment; 9.3% reported past counseling or psychotherapy, psychiatric medication, or both, and (in contrast to the other two community samples) current users of psychotherapy or medication were excluded from participating. Increasingly, social science researchers have used Mturk as a source for research participants, finding that they are more demographically representative and diverse than

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

most samples of convenience, although they tend to skew young, male, and politically liberal (Berinsky, Huber, & Lenz, 2012; Richey & Taylor, 2012). Demographically, 62.7% were male, 76.3% identified as White or Caucasian, 9.5% identified as Asian, 5.6% identified as Black or African American, 5.1% identified as Hispanic or Latino, and 3.5% reported mixed race/ethnicity or "other." All 708 completed the ROS, and 417 of the 708 completed the OQ.

A second community sample (Community2; $n = 102$) was obtained from the adult customers of a full-service car wash. Community2 participants completed a counterbalanced administration of the ROS, ORS, and OQ. Of the 102 participants, 101 provided complete ORS and ROS data, and 99 completed all three measures. Of the 101 participants, 53.5% were female, 82.4% identified as White or Caucasian, 5.9% identified as mixed or multiple races or ethnicities, and 5.0% identified as Latino or Hispanic. Notably, 12.9% of the Community2 sample reported they were currently receiving psychotherapy, psychiatric medication, or both; however, excluding them had no appreciable impact on the descriptive statistics, so they were included in the sample.

A third community sample (Community3; $n = 116$) was drawn from a social and professional network centered in Ireland and the United Kingdom. Originally, 162 people consented to participate, with 116 completing the ROS and ORS in counterbalanced order at all three time points in a Web-based survey at 1-week intervals (actual days between administrations $M = 7.53$, $SD = 1.93$). All 116 identified as White, Caucasian, or European, and 54.3% were female. In the Community3 sample, 7.8% reported that they were currently receiving psychotherapy and/or psychiatric medication; and since excluding them had no appreciable impact on the descriptive statistics, they were included in the sample.

Measures

Rating of Outcome Scale (ROS). The ROS is a three-item measure of well-being using an 11-point Likert-type response format for each item (item range: 0–10; total score range: 0–30). The scale takes less than 30 s to administer and 2–4 s to score by hand. Clients are presented with the measure at the start of each session and instructed to fill in one of the circled numbers from 0 to 10 for each of the three items based on how they "have been feeling in the past week." Instructions and anchors show that a score of 0 indicates the *worst* they have felt and a score of 10 indicates the *best* they have felt.

The three ROS items are modeled on the OQ's three subscales: internal well-being, relational well-being, and task-functioning well-being. However, the ROS items pertain to degree of distress rather than frequency of distressing experiences as in the OQ. The first ROS item (corresponding to the OQ's Symptom Distress subscale) asks, "How do I feel inside, how are my 'gut' feelings?" The wording is in line with the OQ's many felt-sensation-oriented items in this subscale (e.g., "I tire quickly," "I feel weak," and "I have an upset stomach").

The second ROS item, "How are my relationships with people I care about?" was constructed following feedback from the first author's clients using the ORS that a broader array of relationships were of importance to clients' well-being than those specified in the ORS item: "Interpersonally: (Family, close relationships)." The ROS item corresponds well with Interpersonal Relations subscale items on the OQ that involve general relational needs (e.g., "I get along well with others" and "I feel loved and wanted").

The third ROS item, "How am I doing with tasks (work, school, home)?" is more similar to the OQ items about social role (e.g., "I feel stressed at work/school," and "I find my work/school satisfying,") than the ORS item: "Socially: (Work, School, Friendships)." The ROS item was broadened to include "home" to accommodate those who work from home, are stay-at-home parents, or who are retired or otherwise not employed but still have task functions in which they want to engage effectively to experience higher levels of well-being.

Outcome Rating Scale (ORS). The ORS (Miller et al., 2003) is an ultrabrief four-item well-being measure using a 100-mm visual analogue scale response format for each item, and is scored by summing the distances of the four marks from their respective zero points (item range: 0.0–10.0 cm; total score range: 0.0–40.0 cm). It takes less than 1 min to administer and 30–60 s to score by hand using a ruler and calculator (computer-based administration and scoring are available through several publishers). Clients make a mark on each line corresponding to how they "have been doing" in the past week with marks to the left indicating "low levels" and marks to the right indicating "high levels."

Outcome Questionnaire-45.2 (OQ). The OQ consists of 45 items in a 5-point Likert-type response format (item range: 0–4, total score range: 0–180; Lambert et al., 2004). The OQ is a frequency-of-distress scale; therefore higher scores indicate higher distress (i.e., greater frequency of negative experiences, lower frequency of positive experiences) rather than greater well-being. It takes approximately 5 to 10 min to administer and 3–6 min to score by hand. There are three subscales: symptom distress, interpersonal relations, and social role. Subscale items along with interpolated items are summed for each subtotal, and the three subtotals are summed for a total score.

Statistical Analyses, Assumptions, and Benchmarks

To standardize scores between instruments with different ranges and variances, *T* scores were calculated for the ROS, ORS, and OQ, based on available community-sample means and standard deviations. Sources for the ROS were the Community1, Community2, and Community3 samples in this study; sources for the ORS were Community2, Community3, Miller et al. (2003), and Bringhurst et al. (2006) samples; sources for the OQ were Community1, Community2, the first four samples listed in Table 1 of Lambert et al. (2004), Miller et al. (2003), and Bringhurst et al. (2006) samples.

Equivalence between community samples was tested via independent *t* tests (two tailed). Bland-Altman 1-week repeatability (test–retest) plots of ROS and ORS *T* scores were generated for Community3, along with coefficients of repeatability (CR). The CR is conceptually similar to the reliable change index (RCI; see below), and is a measure of the "smallest possible change. . . that represents true/real change. . . [and] accounts for both random and systematic error" (Vaz et al., 2013, p. 6). CR was calculated using both Bland and Altman (1986) and Vaz et al. (2013) methods. In addition to the repeatability plots, Bland-Altman agreement plots between Session-1 *T* scores on the ROS, ORS, and OQ were generated for the clinical sample.

Correlational analysis (including Steiger's *Z* test) between concordant and discordant items was used to explore discriminant validity. Internal consistency was assessed with Cronbach's alpha. Interitem correlation coefficients (Pearson's *r*) were calculated to explore the association between ROS and ORS items and the three OQ subscales. Sensitivity to change was assessed through *t* tests and effect size corrected for repeated measures (ES_{RMC}; Seidel, Miller, & Chow, 2014).

Test–retest reliability (Time 1 to Time 2 in the Community3 sample, Pearson's *r*) was used to calculate RCIs (Jacobson & Truax, 1991). *T* score conversions of RCIs were calculated by dividing each instrument's RCI by its respective raw score standard deviations and multiplying by 10. Raw score clinical cutoffs for the ROS, ORS, and OQ were established via Jacobson and Truax's (1991) criterion *c,* using the clinical sample means and standard deviations ($N = 279$) and the unweighted mean of all available nonclinical sample means and standard deviations (ROS: Community1, Community2, and Community3; ORS: Community2, Community3, Miller et al., 2003, and Bringhurst et al., 2006; OQ: Community1, Community2, Table 1 in the work of Lambert et al., 2004, Miller et al., 2003, and Bringhurst et al., 2006). A common *T* score-based clinical cutoff was set after adjusting for the directionality of measures (i.e., higher *T* scores corresponding with greater well-being).

Finally, previously published clinical cutoffs for the ORS and OQ (i.e., 25 and 63.5, respectively) were compared with the common *T* score cutoff for the ROS, ORS, and OQ. Discordance between measures in categorizing members of the clinical sample ($n = 279$) as distressed or nondistressed was assessed at Time 1 using McNemar's test with Yates's correction for continuity (−0.5). Instrument agreement also was assessed with Cohen's kappa (κ).

Results

T score means and *SD*s of total scores for the clinical and community samples are shown in Table 1. *T* scores for the ROS, ORS, and OQ were based on averaged unweighted community-sample means (22.31, 29.60, and 46.98, respectively) and standard deviations (4.55, 6.80, and 19.40, respectively). The community samples differed from one another, with Community1 most distressed (though still about 1 *SD* better than the clinical sample on the ROS and about 0.5 *SD* better than the clinical sample on the OQ; see Table 1) and Community2 least distressed, expressed in significant differences (all *p*s $< .01$) on all possible independent *t* tests between the samples on the ROS, ORS, and OQ. Community1 and Community2 differed from one another by almost 1 *SD* on both the ROS ($t = 7.173$, $p < .001$, $d = 0.91$) and OQ ($t = 5.716$, $p < .001$, $d = 0.83$). Community1 and Community3 differed from one another by 0.4 *SD* on the ROS ($t = 3.342$, $p < .002$, $d = 0.40$). Community2 and Community3 differed from one another by about 0.5 *SD* on both the ROS ($t = 4.116$, $p < .001$, $d = 0.56$) and ORS ($t = 4.162$, $p < .001$, $d = 0.57$).

Table 1
T Score Means and Differences

Sample	*n*	Test 1 *M*	Test 1 *SD*	Test 2 *M*	Test 2 *SD*	Test 3 *M*	Test 3 *SD*	*t*	*p*
				ROS					
Clinical	279	35.95	12.15						
Clinical (T1–T3)	73	34.29	11.64	40.55	11.86	42.99	11.64	7.29	<.001
Community1	708	45.61	11.85						
Community2	101	54.73	8.77						
Community3	116	49.63	9.38	49.77	9.40	49.65	10.87	.02	.98
(Distressed)	21	35.71	5.67	40.52	10.97	40.21	10.62	2.31	.03
				ORS					
Clinical	279	35.98	11.60						
Clinical (T1–T3)	73	34.39	11.42	39.60	12.40	44.75	13.27	7.31	<.001
Community2	101	53.75	8.63						
Community3	116	48.32	10.34	48.82	11.08	49.55	11.20	1.55	.12
(Distressed)	28	35.31	6.99	40.20	12.06	41.38	11.60	3.31	<.01
				OQ					
Clinical	279	39.57	10.98						
Clinical (T1–T3)	73	38.47	11.37	41.78	11.50	43.38	10.71	6.21	<.001
Community1	417	45.17	13.83						
Community2	99	53.87	8.57						
				Between Instruments					
Clinical ROS × ORS	279	35.95	12.15	35.98	11.60			.07	.94
Clinical ROS × OQ	279	35.95	12.15	39.57	10.98			6.40	<.001
Clinical ORS × OQ	279	35.98	11.60	39.57	10.98			6.73	<.001

Note. Paired-sample, two-tailed *t* tests were conducted between Time 1 and Time 3. Distressed subsample from Community3 only included participants who reported no current psychotherapy or medication. Between-instruments *t* tests were concurrent at Session 1. ROS = Rating of Outcome Scale. ORS = Outcome Rating Scale. OQ = Outcome Questionnaire 45.2. T1–T3 = cohort with three administrations.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Cronbach's alpha and test–retest reliabilities are reported for the ROS and ORS in Table 2. Interitem correlations are shown in Table 3. The ROS alpha levels were good, given only three items (alpha increases as a function of the number of items and item redundancy). The mean interitem correlation was .48 (range = .39–.60) for the clinical sample, and .61 (range = .54–.65) for the community sample. The ROS and ORS had similar test–retest reliabilities.

Bland-Altman 1-week repeatability plots for ROS and ORS *T* scores in Community3 were assessed for bias, following the method suggested by Bland and Altman (1986). There was minimal bias between test and retest for either the ROS or the ORS (calculated by subtracting Time-2 *T* scores from Time-1 *T* scores and determining the mean difference: $M_{diff} \leq 0.5$; Table 2 and Figure 1). However, as expected, the Limits of Agreement (LoA) between test and retest were somewhat broad in this community sample, yielding average CR for the ROS and ORS of 13.7 and 17.0 (i.e., 95% of retest scores falling within 1.37 and 1.70 *SD* of test scores), respectively (see Table 2).

Bland-Altman agreement plots between clinical-sample *T* scores from the ROS, ORS, and OQ at Session 1 were broad as well, often exceeding 1 *SD* of difference between simultaneous scores on different instruments (see Figure 2), and the 95% LoA ranging from 1.5 to over 2 *SD* (see Table 2). Bias between the ROS and ORS was virtually nil, with 49% of the difference scores (i.e., ORS *T* score minus ROS *T* score) being negative ($M_{diff} = 0.03$). Between concurrent ROS and ORS administrations, 97.1% of the clinical sample was within ±20 *T* score points (2 *SD*) of difference, 84.6% was within ±10 *T* score points of difference, and 57.3% was within ±5 points of difference. Bland and Altman's (1986) method of bias adjustment was used on the OQ scores: the average M_{diff} of 3.61 from the between-instruments ROS × OQ and ORS × OQ *T* scores (see Table 2) was subtracted from OQ *T* scores. After adjusting for bias, 39.4% of the sample was within ±5 points of difference ($M_{diff} = 0.15$) between OQ and ROS; and 39.4% of the sample was within ±5 points of difference ($M_{diff} = 0.11$) between OQ and ORS.

The results of *t* tests and the LoA between measures at Session 1 are provided in Tables 1 and 2, respectively. Clinical-sample ROS and ORS scores were (as a group) quite similar to each other; but both instruments yielded significantly lower well-being *T* scores at Session 1 than the OQ did (via *t* test; see Table 1).

Discriminant validity for an ultrabrief scale of three items cannot be assessed through traditional methods (e.g., Campbell & Fiske, 1959). However, preliminary support for the ROS as a measure of psychological well-being distinct from a measure of

Table 2
Reliability, Agreement, and Change Indices for T *scores*

Sample	*n*	α	r_{xx}	M_{diff} (bias)[a]	SD_{diff}[a]	ES_{RMC}[b]	95% LoA LB (95% CI)	95% LoA UB (95% CI)	CR	RCI
						ROS				
Clinical	279	.73								
Clinical (T1-T3)	73	.72		6.26	11.94	.75				
Community1	708	.80								
Community2	101	.74								
Community3	116	.76	.73[a]	.13	6.90	.00	−13.7 (−15.9 to −11.5)	13.4 (11.2 to 15.6)	13.8, 19.2	14.5
(Distressed)	21					.48				
						ORS				
Clinical	279	.81								
Clinical (T1–T3)	73	.84		5.21	11.48	.83				
Community2	101	.87								
Community3	116	.86	.68[a]	.50	8.57	.11	−17.3 (−20.0 to −14.6)	16.3 (13.5 to 19.0)	17.1, 23.8	15.7
(Distressed)	28					.59				
						OQ				
Clinical	279									
Clinical (T1–T3)	73			3.31	6.54	.44				
Community1	417									
Community2	99									
						Between instruments				
Clinical ROS × ORS	279		.80[c]	.03	7.60		−15.2 (−15.9 to −14.4)	15.2 (14.5 to 16.0)		
Clinical ROS × OQ	279		.67[c]	3.63	10.11		−16.6 (−17.6 to −15.6)	23.8 (22.8 to 24.8)		
Clinical ORS × OQ	279		.69[c]	3.59	9.60		−15.6 (−16.6 to −14.7)	22.8 (21.8 to 23.7)		

Note. Distressed subsample from Community3 only included participants who reported no current psychotherapy or medication. α = Cronbach's alpha. r_{xx} = test–retest correlation; SD_{diff} is between subjects; ES_{RMC} = repeated-measures corrected effect size; LoA LB = limits of agreement, lower bound; LoA UB = limits of agreement, upper bound; CI = confidence interval; CR = coefficient of repeatability (formulas from Bland & Altman, 1986; Vaz et al., 2013); RCI = reliable change index; ROS = Rating of Outcome Scale. ORS = Outcome Rating Scale; OQ = Outcome Questionnaire 45.2; T1–T3 = cohort with three administrations.
[a] Time 1 to Time 2. [b] Time 1 to Time 3. [c] Concurrent reliability.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Correlations Between Well-Being Components at Time 1

	OQ *SD*	OQ IR	OQ SR	ROS 1	ROS 2	ROS 3
OQ *SD*	1	*.73*	*.79*	**−.68**	*−.60*	*−.67*
OQ IR	.48	1	*.68*	*−.58*	**−.74**	*−.59*
OQ SR	.70	.42	1	*−.56*	*−.55*	**−.65**
ROS 1	**−.60**	−.38	−.40	1	*.65*	*.64*
ROS 2	−.30	**−.50**	−.25	.44	1	*.54*
ROS 3	−.61	−.37	**−.59**	.60	.39	1
	OQ *SD*	OQ IR	OQ SR	ORS 1	ORS 2	ORS 3
OQ *SD*	1	.54	.58	**−.53**	−.47	−.54
OQ IR	.48	1	.48	−.40	**−.63**	−.52
OQ SR	.70	.42	1	−.37	−.31	**−.32**
ORS 1	**−.58**	−.28	−.42	1	.55	.57
ORS 2	−.27	**−.48**	−.21	.36	1	.67
ORS 3	−.58	−.38	**−.59**	.54	.29	1

Note. Community1 + 2 (values in italics; $n = 516$) and Community2 ($n = 99$) sample correlations are above the diagonals. Clinical sample ($n = 279$) correlations are below the diagonals. Corresponding items and subscales are in bold type. Pearson's r correlations between OQ and the other instruments are negative due to reverse scoring of the OQ (higher OQ scores indicate distress rather than well-being). OQ = Outcome Questionnaire 45.2; SD = Symptom Distress Subscale; IR = Interpersonal Relationships subscale; SR = Social Role subscale; ROS = Rating of Outcome Scale; ORS = Outcome Rating Scale; ROS 1 = ROS item 1. Two tailed, all significant at $p < .01$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

physical well-being (a related but different construct) was provided by correlational analysis between ROS total score and a selection of three OQ items in the clinical ($n = 279$) and Community1 ($n = 413$) samples. Two psychological well-being items from the OQ were reported by Lambert et al. (2004) as discriminating well between clients and controls in terms of change from Time 1 to Time 2 ("I feel fearful" and "I feel blue"). One physical well-being item ("I have sore muscles") was reported by Lambert et al. (2004) as not discriminating well between clients and controls. Correlations between the OQ items and the ROS total score were .39 and .53 for the concordant items ("I feel fearful" and "I feel blue") and .19 for the discordant item ("I have sore muscles") in the clinical sample. Steiger's Z was <2.74 ($df = 274$, $p < .01$) between the correlation of ROS and each OQ concordant item versus the correlation of ROS and the OQ discordant item. Correlations between the OQ items and the ROS total score were .60 and .68 for the concordant items and .21 for the discordant item in the Community1 sample. Steiger's Z was <7.34 ($df = 411$, $p < .01$) between the correlation of ROS and each OQ concordant item versus the correlation of ROS and the OQ discordant item. These significant differences provide some support for the ROS as a measure of psychological well-being.

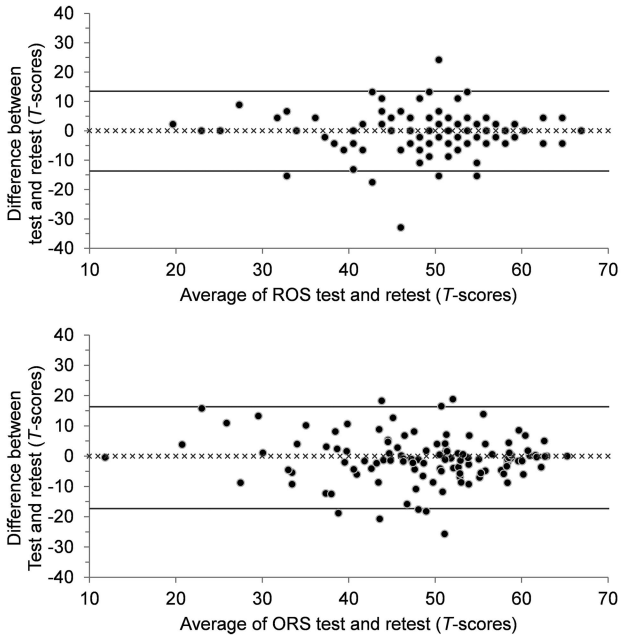

Figure 1. Bland-Altman repeatability plots for the ROS and ORS (Community3 sample). Lines represent the 95% confidence interval of the limits of agreement. ROS = Rating of Outcome Scale; ORS = Outcome Rating Scale.

To illustrate the association between elements of the well-being construct on the different measures, correlations between ROS and ORS items and OQ subscales for both the community and clinical samples are presented in Table 3. Correlations between similar constructs (e.g., ROS item 1 and OQ Symptom Distress) are in bold type, with the ROS and OQ items yielding correlation coefficients between .50 and .60 in the clinical sample, and between .65 and .74 in the community sample. Correlations between less identical aspects of well-being (e.g., task-oriented and relationship-oriented items) were generally, but not consistently, lower.

The repeated-measures cohort ($n = 73$) of the clinical sample reported about 0.4 *SD* more distress on the ROS and ORS at intake (Session 1) than on the OQ. By the third session, these differences between measures diminished. However, despite the convergence between the different measures by Session 3, Figure 3 shows the effect of the Session-1 difference on repeated-measures-corrected pre-post estimates of change (ES_{RMC}) from Session 1 to Sessions 2 and 3. The Sessions 1–3 ES_{RMC} for the ROS, ORS, and OQ were 0.75, 0.83, and 0.44, respectively (Figure 3 and Table 2). The

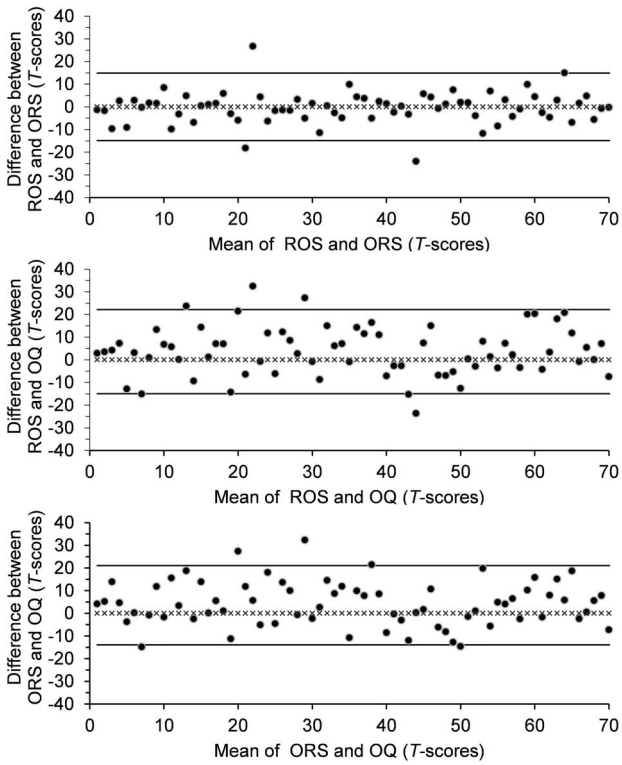

Figure 2. Bland-Altman agreement plots for the ROS, ORS, and OQ (clinical sample). Lines represent the 95% confidence interval of the limits of agreement. ROS = Rating of Outcome Scale; ORS = Outcome Rating Scale; OQ = Outcome Questionnaire.

results of paired *t* tests between Session 1 and 3 were similarly significant for the three measures (see Table 1). Figure 3 shows relatively little change on the ROS and ORS for Community3 from Time 1 to Time 2 and 3; however, the distressed (but untreated) subsample of Community3 showed considerable clinical change (ES_{RMC} = 0.48 and 0.59, respectively) by Time 3.

RCIs were calculated as critical values based on Jacobson and Truax's (1991) method. That is, the standard error of the difference (S_{diff}) based on test–retest reliability and standard deviation of Time 1 scores averaged across community samples was multiplied by 1.96 (i.e., $p = .05$, two tailed). RCIs (i.e., critical values) for the ROS, ORS, and OQ were 6.6, 10.7, and 18.6 points, respectively; *T* score RCIs were 14.5, 15.7, and 9.6. These newly calculated RCIs reduced the variability between how these instruments measured clinical change and also reduced the proportion of people classified as changed (see Figure 4). Raw score clinical cutoffs for the ROS, ORS, and OQ's "distressed range" scores were calculated as ≤19.4, ≤ 25.2, and ≥56.6, respectively. At these cutoff scores, an average of 20.5% of the clinical sample scored in the nondistressed range on both of two instruments at Session 1 (21.5% on both ORS and OQ, κ = .49; 19.4% on ROS and OQ, κ = .44; and 20.8% on ORS and ROS, κ = .55). As a cross-method check on the raw score cut-off for the ROS, a Response Operator Characteristics (ROC) analysis was undertaken, using the clinical sample (N = 279) versus all community samples (N = 925). The area under the curve (AUC) was .75 ($p < .001$; 95% confidence interval [CI] [.72, .78]). The ROC cutoff was 19.5 for the ROS at the point of balance between sensitivity and specificity (.71 and .66, respectively). Using all available community samples for the ORS (n = 217) and OQ (n = 516) raw scores yielded ROC-based cutoffs of 26.5, and 58.5, respectively.

Jacobson and Truax's (1991) criterion *c* formula was applied to the *T* scores (with high scores corresponding to greater well-being) yielding cutoff values within 0.5 points of 44.0 for each measure. The *T* score value of 44 categorized a similar proportion as nondistressed (namely, 19.7% of the clinical sample scoring in the nondistressed range on both of two instruments at Session 1: 20.4% on ORS and OQ; 19.7% on ROS and OQ; and 19.0% on ORS and ROS). In comparison, using a *T* score cutoff of 43 led to an average of 15.9% scoring as nondistressed on both of two instruments, and a *T* score cutoff of 45 led to an average of 22.3% of the sample being characterized as nondistressed). The cutoff based on an approximately 20% nondistressed subsample established a common best-fit cutoff *T* score of <44 for clinical distress. As a cross-method check on the *T* score cut-off for the ROS, a ROC analysis was performed using the clinical sample (n = 279) versus all community samples (n = 925), leading to the same results as from the raw scores: The AUC was .75 ($p < .001$; 95% CI [.72, .78]). The ROC cutoff was 44 for the ROS at the point of balance between sensitivity and specificity (.71 and .66, respectively). Using all available community samples for the ORS (n = 217) and OQ (n = 516) raw scores yielded ROC-based *T* score cutoffs of 45.5, and 44.1, respectively.

Using the *T* score RCIs and clinical cutoff resulted in 70% agreement between the OQ and ORS and 71% agreement between the OQ and ROS in reliable change classifications. For clinically significant change classifications, the OQ and ORS agreed in 74% of cases, as did the OQ and ROS. In contrast, previously published RCIs and cutoffs resulted in a 63% agreement between the OQ and ORS for both reliable change and clinically significant change classifications.

Using the cutoff score of <44 for the clinical sample (n = 279), McNemar's test showed no discordance between ROS and ORS *T*

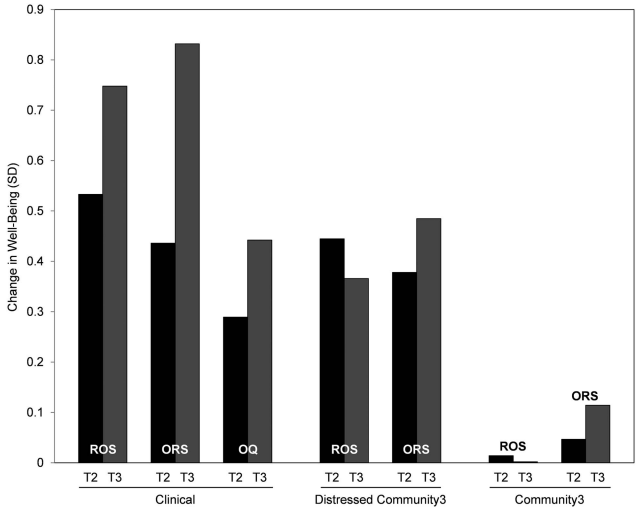

Figure 3. Change in well-being *T* scores from Time 1 to Time 2 and 3 (clinical and Community3 samples). ROS = Rating of Outcome Scale; ORS = Outcome Rating Scale; OQ = Outcome Questionnaire.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

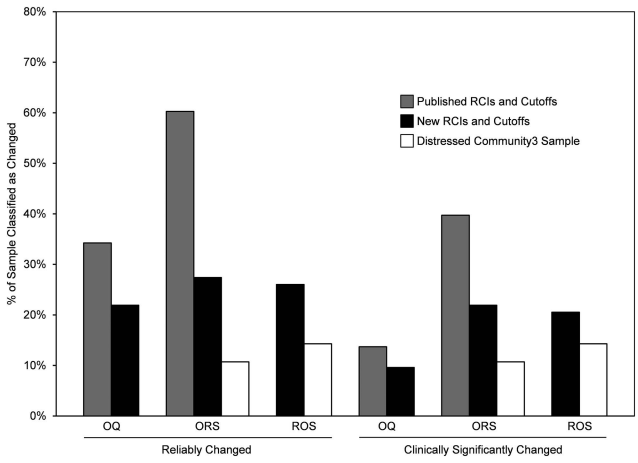

Figure 4. Effect of new RCIs and clinical cutoffs on change classifications in the clinical sample and distressed Community3 subsample (Time 1 to Time 3). ROS = Rating of Outcome Scale; ORS = Outcome Rating Scale; OQ = Outcome Questionnaire; RCI = reliable change index.

score-based categorization of "distressed" and "nondistressed" ($\chi^2 = 0.005$, $df = 1$, $p = .95$, $\kappa = .52$). However, there was discordance from the OQ *T* scores on both the ROS ($\chi^2 = 5.28$, $p = .02$, $\kappa = .42$) and the ORS ($\chi^2 = 6.27$, $p = .01$, $\kappa = .46$), with both of the ultrabrief measures classifying more people as distressed than the OQ did.

Discussion

This study addressed the equivalence of two ultrabrief measures of general well-being (both based on the OQ) for assessing clinical outcomes. Cronbach's alpha and test–retest reliabilities for the new ROS were good, especially for a three-item measure. Mean interitem correlations and ranges were also good (Clark & Watson, 1995). Bland-Altman analysis showed virtually no bias between the ROS and ORS. Parameters for acceptable agreement have not been established for method-comparison studies of well-being, and agreement is likely to be modest given the complexity and variability of a brief, general, psychological well-being construct. Using a rather stringent benchmark of ± 0.5 *SD* of difference as "agreement" for comparative purposes, the 57% agreement between the ultrabrief measures versus the 39% agreement between each of these measures and the OQ exemplifies the degree to which the measures vary (using a very broad benchmark of ± 2 *SD* yielded >96% agreement for all three comparisons).

The ROS's sensitivity to change was excellent in a clinical sample that was expected to change, and appropriately low in a community sample that was not expected to change, however both the ROS and ORS Session-1 scores showed more distress than did the OQ scores, indicating that these ultrabrief measures may be more sensitive to change due to clients' more distressed responses (relative to community norms) prior to treatment. Correlations between the ROS items and OQ subscales compared favorably with those between the ORS and OQ. Discriminant validity for the ROS was explored by comparing the total score of this general well-being scale with two psychologically oriented items and one somatic item on the OQ with differential responses to psychotherapy in previous research. The ROS correlated better with the psychological items than the somatic item. These findings together suggest good sensitivity, specificity, and construct validity.

RCIs and clinical cutoffs were calculated and compared with previously published benchmarks and methods, and the new methods improved concordance between measures, as illustrated in Figure 4. Results from ROC analyses very closely replicated the Jacobson & Truax-based methods for raw and *T* score-based clinical cutoffs of 19.4 and 44 for the ROS, and showed good diagnostic accuracy based on the AUC. Using Jacobson and Truax's (1991) recommendation of the test–retest reliability coefficient (for a community sample not receiving a clinical intervention) with an unweighted mean of independent community-sample standard deviations, resulted in a more stringent categorization of reliable change compared with previous methods (e.g., Lambert et al., 2004; Miller et al., 2004). The current method also reduced the variability between the different instruments in how they categorized change in contrast to previously reported RCIs (that were based on Cronbach's alpha). In contrast to an RCI of 5 calculated by Miller and Duncan (2004), we obtained a substantially greater raw score RCI of 10.7. In comparison, Janse et al. (2014) found an RCI of 9 for the ORS in their Dutch sample (no details were provided about their method). It may be that different sample variances and different RCI construction methods may contribute to the variability in the RCI for the ORS among different studies. We recommend Jacobson and Truax's original method of test–retest reliability (rather than Cronbach's alpha) as the most appropriate basis for determining measurement error. In the context of judging the reliability of scores from an instrument purportedly measuring clinical change, it is most important to account for error in the form of moment-by-moment fluctuations in well-being that everyone may experience from time to time, such that clinically important fluctuations of well-being are those beyond these "background" levels. Establishing the reliability in a relatively short time (1 week) in an untreated community sample provides the appropriate background levels of change for estimating reliable clinical change from these scores. It should be noted that slightly less change was required on the ROS than the ORS to meet CR and RCI thresholds for reliable or meaningful change. Thus, while the ORS appeared to be more sensitive to change (as evidenced in greater effect sizes), the ROS provided a modest improvement in the "signal-to-noise ratio" between change and error.

The new, *T* score-based clinical cutoff of <44 reduced the differences between scales in the categorization of clients as clinically significantly changed (CSC; i.e., initially distressed clients with change ≥ RCI and crossing the clinical cutoff), although substantial differences remained in CSC classifications between the OQ and the ultrabrief measures. It should be noted, in addition, that much of the gain in agreement in CSC was due to fewer respondents meeting the new RCI and cutoff thresholds. Despite the improvements in agreement between the three measures by using these new methods, significant discordance remained between the ultrabrief measures and the OQ in terms of classifying people as distressed or not distressed in Session 1.

Finally, the degree to which distressed community members showed significant improvement from test to retest deserves comment. While at first this might cause concern about the degree to which change in a clinical sample should be considered a result of treatment (if community members can change so dramatically

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

without it), the measures do not differentiate between people in terms of how long they have suffered prior to assessment. People generally do not seek treatment for transitory disturbances in their well-being, but rather when their efforts to improve it have not been effective. Therefore, the improvement of the distressed community subsample is not comparable to that which might be seen in an untreated or waitlisted clinical sample.

The current study is limited in several ways that need to be addressed in future research. The variation among independently sampled communities of adults shows the importance of recruiting a variety of independent samples using different sampling methods to better represent the community norms against which one wishes to benchmark well-being and change. Community samples are not the same as "nonclinical samples." All communities contain a proportion of clinically distressed individuals who may or may not be in treatment. Normative samples may not obtain data about current treatment (e.g., Lambert et al., 2004). In the current study, the community samples were questioned about current treatment, however inclusion of these data was not uniform between the samples (i.e., Community1 excluded current treatment recipients, while the other two community samples included these participants). While highly practical, using only ultrabrief outcome instruments in clinical practice limits the information practitioners and clients can use to make informed clinical decisions and judge clinical change. One way to balance the desire for more detailed information with the feasibility of ultrabrief session-by-session measurement is to use longer instruments intermittently: for example, at every fifth session. However, the apparently greater sensitivity of the simultaneously administered ultrabrief instruments versus the OQ leaves open a different question: to what extent are different instruments that purport to measure general subjective well-being interchangeable? It is unclear, for example, whether the apparently greater sensitivity of the ultrabrief instruments to distress at intake was the result of the length of these instruments or the nature of the items (which—while based on the OQ subscales—were different from the longer OQ in the generality of the item wording, the greater range of their item-response formats, and the focus on degree of distress rather than frequency of experience). These questions clearly require a greater research focus on the cross-validation of instruments used to measure psychotherapy outcomes.

Other limitations are the use of a single clinical sample for estimating the relative sensitivity-to-change among instruments; and the lack of diagnostic categorizations of clients or standardized administration or exclusion of medication (though it should be noted that distress levels at intake were comparable with those reported in clinical samples from previous research, e.g., Anker, Duncan, & Sparks, 2009; Bauer, Lambert, & Nielsen, 2004; Miller et al., 2003; Okiishi et al., 2006; Reese et al., 2009). Further research in diverse settings will clarify the extent to which the current findings can be generalized to other client samples. A related limitation was the small number of therapists in the sample which precluded nested analyses that might further clarify the relationship between specific instruments, therapist factors (such as the way individual therapists might administer the instruments), and measured change. The first item of the ROS ("How have I felt inside, how have my 'gut' feelings been?") implies both interoceptive and intuitive appraisals of experience. These related processes have been shown to influence motivation, awareness, and interpersonal behavior (Dunn et al., 2010; Farb et al., 2015; Garfinkel & Critchley, 2013). Yet, the wording of the item may not direct clients clearly enough to focus their attention on interoception. Further development of this item (particularly in the process of translation to other languages and cultures) is warranted. Finally, this study examined the degree of change over three administrations. It is possible that differences between measures might attenuate or intensify through a longer course of treatment, and that different conclusions might be drawn if longer-term data were collected.

There has been increasing focus on greater accountability from clinicians in monitoring clients' treatment progress; and there is compelling evidence that formally monitoring progress can improve clinical outcomes. However, the length of most measures and the lack of a clear framework for making sense of outcome data are hurdles preventing greater utilization. For clinician-researchers who use PROMs in FIT or other forms of practice-based evidence, the ROS appears to be a good instrument to consider and it adheres to guidelines set forth by the National Quality Forum: it measures important, face valid aspects of well-being to which clients relate; it shows promising levels of reliability and validity; it is extremely feasible to administer; and it demonstrates appropriate responsiveness to clinical change without being overly reactive. By using rigorous but easy-to-calculate common benchmarks for interinstrument crosswalks, its data appear to be comparable (with some caveats) to two of the most broadly available outcome measures currently in use: the ORS and the OQ. More importantly, methods for assessing equivalence between instruments require continued exploration to increase the validity of score interpretations from ultrabrief PROMs.

References

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology, 77,* 693–704. http://dx.doi.org/10.1037/a0016062

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61,* 271–285. http://dx.doi.org/10.1037/0003-066X.61.4.271

Australian Mental Health Outcomes and Classification Network. (2005). *Adult national outcomes & casemix collection standard reports* (1st ed., version 1.1). Brisbane, Australia: Author.

Barkham, M., Bewick, B. M., Mullin, T., Gilbody, S., Connell, J., Cahill, J., . . . Evans, C. (2013). The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. *Counselling & Psychotherapy Research, 13,* 3–13. http://dx.doi.org/10.1080/14733145.2012.729069

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment, 82,* 60–70. http://dx.doi.org/10.1207/s15327752jpa8201_11

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20,* 351–368. http://dx.doi.org/10.1093/pan/mpr057

Bertolino, B., & Miller, S. D. (Eds.). (2013). *ICCE manuals on feedback-informed treatment* (Vols. 1–6). Chicago, IL: ICCE Press.

Bijur, P. E., Silver, W., & Gallagher, E. J. (2001). Reliability of the visual analog scale for measurement of acute pain. *Academic Emergency Medicine, 8,* 1153–1157. http://dx.doi.org/10.1111/j.1553-2712.2001.tb01132.x

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 327,* 307–310. http://dx.doi.org/10.1016/S0140-6736(86)90837-8

Bringhurst, D. L., Watson, C. S., Miller, S. D., & Duncan, B. L. (2006). The reliability and validity of the outcome rating scale: A replication study of a brief clinical measure. *Journal of Brief Therapy, 5,* 23–29.

Campbell, A., & Hemsley, S. (2009). Outcome rating scale and session rating scale in psychological practice: Clinical utility of ultra-brief measures. *Clinical Psychologist, 13,* 1–9. http://dx.doi.org/10.1080/13284200802676391

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

Connell, J., & Barkham, M. (2007). *CORE-10 user manual (version 1.1)*. Rugby, UK: CORE System Trust and CORE Information Management Systems Ltd.

DeLoach, L. J., Higgins, M. S., Caplan, A. B., & Stiff, J. L. (1998). The visual analog scale in the immediate postoperative period: Intrasubject variability and correlation with a numeric scale. *Anesthesia and Analgesia, 86,* 102–106.

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., . . . Dalgleish, T. (2010). Listening to your heart. How interoception shapes emotion experience and intuitive decision making. *Psychological Science, 21,* 1835–1844. http://dx.doi.org/10.1177/0956797610389191

Farb, N., Daubenmier, J., Price, C. J., Gard, T., Kerr, C., Dunn, B. D., . . . Mehling, W. E. (2015). Interoception, contemplative practice, and health. *Frontiers in Psychology, 6,* 1–26. http://dx.doi.org/10.3389/fpsyg.2015.00763

Garfinkel, S. N., & Critchley, H. D. (2013). Interoception, emotion and brain: New insights link internal physiology to social behaviour. Commentary on: "Anterior insular cortex mediates bodily sensibility and social anxiety" by Terasawa et al. (2012). *Social Cognitive and Affective Neuroscience, 8,* 231–234. http://dx.doi.org/10.1093/scan/nss140

Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist–client feedback and clinical support tools. *Psychotherapy Research, 17,* 379–392. http://dx.doi.org/10.1080/10503300600702331

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. http://dx.doi.org/10.1037/0022-006X.59.1.12

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

Kopta, S. M., & Lowry, J. L. (2002). Psychometric evaluation of the Behavioral Health Questionnaire—20: A brief instrument for assessing global mental health and the three phases of psychotherapy outcome. *Psychotherapy Research, 12,* 413–426. http://dx.doi.org/10.1093/ptr/12.4.413

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., & Reid, R. C. (2004). *Administration and scoring manual for the Outcome Questionnaire 45.2*. Salt Lake City, UT: OQ Measures.

Lima, E. P., Barreto, S. M., & Assunção, A. A. (2012). Factor structure, internal consistency and reliability of the Posttraumatic Stress Disorder Checklist (PCL): An exploratory study. *Trends in Psychiatry and Psychotherapy, 34,* 215–222. http://dx.doi.org/10.1590/S2237-60892012000400007

Miller, S. D., & Duncan, B. L. (2004). *The Outcome and Session Rating Scales: Administration and scoring manual*. Chicago, IL: Institute for the Study of Therapeutic Change.

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

Müller, M. C., Cross, N. C. P., Erben, P., Schenk, T., Hanfstein, B., Ernst, T., . . . Hochhaus, A. (2009). Harmonization of molecular monitoring of CML therapy in Europe. *Leukemia, 23,* 1957–1963. http://dx.doi.org/10.1038/leu.2009.168

National Quality Forum. (2013). *Patient reported outcomes (PROs) in performance measurement*. Washington DC: Author. Retrieved March 16, 2015, from http://www.qualityforum.org/Publications/2012/12/Patient-Reported_Outcomes_in_Performance_Measurement.aspx

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielsen, L., Dayton, D. D., & Vermeersch, D. A. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology, 62,* 1157–1172. http://dx.doi.org/10.1002/jclp.20272

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy: Theory, Research, Practice, Training, 46,* 418–431. http://dx.doi.org/10.1037/a0017901

Richey, S., & Taylor, B. (2012, December 19). *How representative are Amazon Mechanical Turk workers?* Retrieved May 25, 2014, from http://themonkeycage.org/2012/12/19/how-representative-are-amazon-mechanical-turk-workers/

Seidel, J. A. (2011). *Rating of Outcome and Session Experience Scales (ROSES)*. (version 2.0). Denver, CO: Author. Available online at http://coloradopsychology.com

Seidel, J. A., Miller, S. D., & Chow, D. L. (2014). Effect size calculations for the clinician: Methods and comparability. *Psychotherapy Research, 24,* 470–484. http://dx.doi.org/10.1080/10503307.2013.840812

Swift, J. K., Greenberg, R. P., Whipple, J. L., & Kominiak, N. (2012). Practice recommendations for reducing premature termination in therapy. *Professional Psychology: Research and Practice, 43,* 379–387. http://dx.doi.org/10.1037/a0028291

Tarescavage, A. M., & Ben-Porath, Y. S. (2014). Psychotherapeutic outcomes measures: A critical review for practitioners. *Journal of Clinical Psychology, 70,* 808–830. http://dx.doi.org/10.1002/jclp.22080

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability. *PLoS ONE, 8*(9), e73990. http://dx.doi.org/10.1371/journal.pone.0073990

Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *Journal of Voice, 13,* 508–517. http://dx.doi.org/10.1016/S0892-1997(99)80006-X

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

(*Appendix follows*)

Appendix

Rating of Outcome Scale (Version 2.1)

Please mark how you have been feeling in the past week, with "0" as the worst and "10" as the best. Please fill in the appropriate circle like this: ●

	Worst										Best
How have I felt inside, how have my "gut" feelings been?	⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
How were my relationships with people I care about?	⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
How have I been doing with tasks (work, school, home)?	⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩

Comments:.....

Copyright 2011. Jason A. Seidel, Psy.D.

Received August 14, 2015
Revision received February 18, 2016
Accepted February 25, 2016 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.