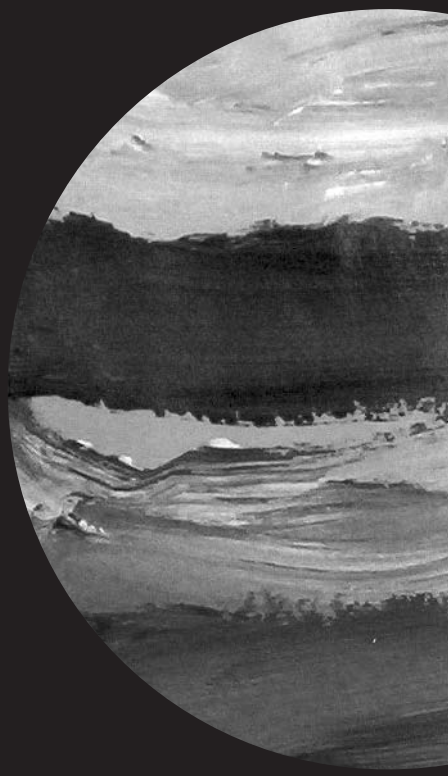
Rewiring efficacy studies to increase their relevance to routine practice

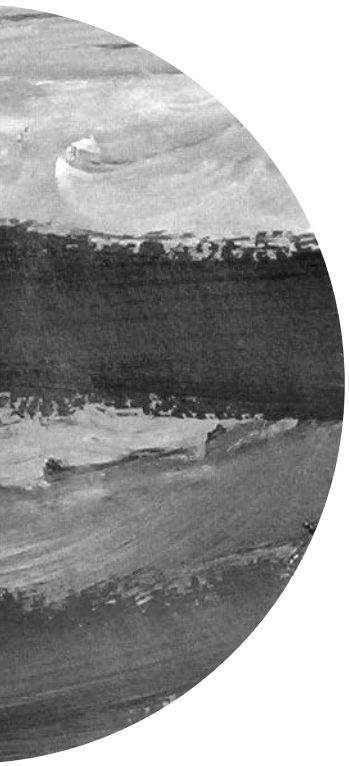


Michael Barkham[1], Chris Leach[2], David A Shapiro[3], Gillian E Hardy[3], Mike Lucock[2] & Anne Rees[1]

[1] Psychological Therapies Research Centre, University of Leeds
[2] South West Yorkshire Mental Health NHS Trust and University of Huddersfield
[3] Universities of Leeds and Sheffield

Rewiring efficacy studies to increase their relevance to routine practice

M Barkham, C Leach, D A Shapiro, G E Hardy, M Lucock & A Rees



Abstract

Current efficacy literature relies heavily on the Beck Depression Inventory (BDI) as the gold standard patient self-report measure. In contrast, the evaluation of psychological therapies in routine practice relies heavily on the CORE-OM. Although the two measures are conceptually distinct, they have been shown to be highly correlated. This suggests the possibility of replacing one measure with the other - a procedure we refer to as rewiring - in service of making the results of efficacy studies using the BDI have greater relevance of practitioners who routinely use the CORE-OM. We tested this proposition using transformation tables (Leach et al., in press) to convert BDI-I scores into CORE-OM scores and reran the analysis of a major efficacy study of depression - the Second Sheffield Psychotherapy Project (Shapiro et al., 1994). Results showed a near perfect replication of the original results and examples of benchmarks concerning the overall effects of treatment as well as differences between treatments are provided against which outcomes in routine practice can be contrasted. The implications for bridging efficacy and effectiveness research are discussed.

Keywords: CORE-OM; Beck Depression Inventory; efficacy; effectiveness; evidence based practice; practice based evidence

Introduction

The gap between research and practice in the area of the psychological therapies has been a continuing theme in the literature (e.g., Chawalisz, 2003). Traditional research has been built on evidence derived from efficacy studies (i.e., randomised or comparative trials) and has culminated in the paradigm of evidence-based practice. By contrast, research activity which is often seen as more relevant to practitioners is built on evidence from studies of the effectiveness of psychological therapies in routine settings (effectiveness studies) and has yielded the paradigm of practice-based evidence (see Barkham & Mellor-Clark, 2000). Rather than seeing these two approaches as competing, it has been argued elsewhere that they are complementary and that both paradigms are needed in order to build a more robust knowledge base and to help bridge the gap between research and practice (see Barkham & Mellor-Clark, 2003). For practitioners, the questions asked within efficacy studies are often not seen to be relevant to them. In other instances, the questions being asked in efficacy trials are relevant but the results cannot be directly transformed into clinical practice because of, for example, the use of different measures and the sampling of different client characteristics. Given this potential mismatch, any developments that facilitate comparisons between efficacy and effectiveness research will help bridge the gap between research and routine practice.

Accordingly, the purpose of this paper is to test a procedure which enables results from efficacy trials to be transformed from the original measure used into one which is widely used in routine clinical practice using formulae based on a large clinical sample and thereby provide a bridge towards making previous trials more relevant to routine services and everyday practitioners. This is a procedure we term rewiring - the simple idea being to replace the old measure with a new one such that the findings can use a currency similar to that used in routine practice.

Attempts have been made to provide some common language for the use of outcome measures in the psychological therapies, most notably the attempt to develop a core outcome battery some 30 years ago (see Waskow, 1975). However, this attempt was aimed solely at trying to identify a common set of measures within the research community and failed. Since then, on the one hand there has been a profusion of outcome measures (see Froyd et al., 1996) but also a freezing of measure development because of the continuing adherence within the research community to key outcome measures (see Horowitz et al., 1997). Adherence to, for example, the

Beck Depression Inventory in its original (BDI-I; Beck et al., 1961) or revised form (BDI-II; Beck et al., 1996) has occurred because of the strongly held view that successive research studies need to use the same measure in order to make comparisons with the existing literature.

In contrast to the use of proprietary measures in efficacy studies, which carry with them a considerable cost burden when used in routine service settings, the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM: Barkham et al., 2001, 2005; Evans et al., 2002; Leach et al., 2004) has become a widely used outcome measure in NHS services. However, while efficacy research continues to use measures such as the BDI and routine practice uses measures such as the CORE-OM, this has the potential for continuing the divide between these research endeavours. If it could be shown that direct comparisons can indeed be made between the existing literature using, for example, the BDI, and newer studies using a different measure, then this is likely to both enhance the use of older literature and reduce the likelihood of a single measure freezing the field.

One recent development had reported procedures for transforming between scores obtained on the BDI-I and the CORE-OM with a high degree of accuracy (Leach et al., in press). This offers the opportunity to test whether the CORE-OM, a generic measure drawn from a pan-theoretical framework, can complement the BDI-I. The original report on the psychometric properties of the CORE-OM found a correlation of .86 with the BDI-I on a sample of patients completing both instruments (Evans et al., 2002). Leach et al. (in press) found an identical correlation between the BDI-I and CORE-OM for clients completing both measures prior to therapy and used their large data set (N = 2234) to devise gender-specific transformation tables for converting BDI-I scores to CORE-OM scores and vice versa.

While Leach et al. (in press) have provided evidence for transformation of scores within effectiveness research (i.e., within routine practice settings), we sought to test the transformation in an archived efficacy study of depression. Working at the interface between efficacy and effectiveness paradigms, we view the BDI versions and CORE-OM as exemplary efficacy and effectiveness measures respectively, whose convergence or divergence needs to be established in a range of different settings.

The aim of this paper was to test the proposition, given the high correlation between BDI-I and CORE-OM scores, that BDI-I scores can indeed be transformed into CORE-OM scores and yield equivalent results in the context of a previously carried out efficacy trial. To achieve this, we used archived data from the Second Sheffield Psychotherapy Project (SPP2; Shapiro et al., 1994, 1995) and kept all parameters of the original study other than substituting the BDI-I with the CORE-OM using the tables derived by Leach et al. (in press).

Method

Data set

The data set comprised the Second Sheffield Psychotherapy Project comparing cognitive behaviour therapy (CB) with Psychodynamic Interpersonal therapy (PI). A total of 117 patients had completed the BDI at 5 time points: screening, intake assessment (A1), session 1 (A2), end of treatment (A3), 3-month follow-up (A4), and 1-year follow-up (A5). Full details are reported elsewhere (see Shapiro et al., 1994). All BDI scores were transformed using the appropriate male or female tables available in Leach et al. (in press). These tables were constructed from transformations based on a combination of non-linear smoothing techniques and non-linear regression, but a good approximation to the transformation tables can be obtained from using the following non-linear regression equations alone: Females: CORE = 0.309 x BDI-I$^{0.60}$ - 0.152; Males: CORE = 0.319 x BDI-I$^{0.60}$ - 0.142. In the analyses reported here, we used the transformation tables for greater accuracy.

Scoring

Following Leach et al. (in press), we aimed at enhancing the clinical meaning of CORE-OM scores to practitioners. Rather than working with the scale 0 to 4, we multiplied the mean item scores by 10 such that the range of human distress fell on

a scale from 0 to 40. We termed this the clinical score. This decision was based on feedback from practitioners stating that they found it easier to assign meaning to a score using a 40 as opposed to a 4-point scale. This procedure does not affect the psychometric properties of the scale.

Analyses: Adjusted scores and covariates

All procedures carried out in the Shapiro et al. (1994) statistical analyses were first replicated on the BDI data alone to ensure the closest match between procedures and SPSS versions used in the analysis of SPP2 data in 1992-3. The procedures were carried out by the same person (AR) who undertook much of the analysis of the SPP2 data set for the Shapiro et al. (1994) publication. Identical procedures were then applied to the CORE-OM score.

In parallel with the analysis used for the BDI in SPP2 (Shapiro et al., 1994), we partialled out Assessment 1 CORE-OM scores. We standardised the Assessment 1 score within severity groups, before entering it as a covariate, to eliminate confounding of the Assessment 1 covariate with the severity factor. To adjust for any mean differences in effectiveness amongst therapists, we used residual scores obtained by subtracting from each adjusted score the mean adjusted score obtained on that occasion by all clients seen by that therapist. All adjusted means reported below are adjusted as described here.

Results

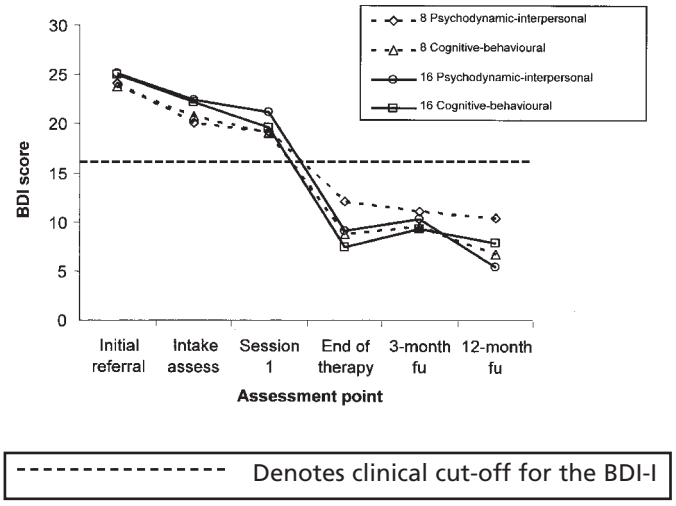


Figure 1: Original BDI-I scores for treatment conditions across treatment and follow-up.

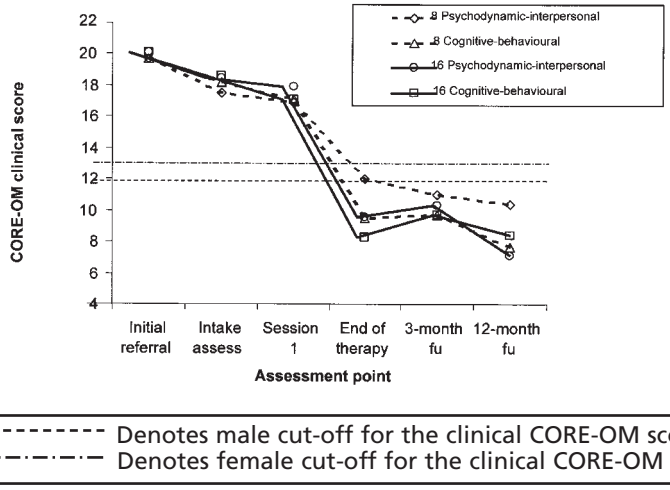


Figure 2: Transformed BDI-I into CORE-OM clinical scores for treatment conditions across treatment and follow-up.

Overall treatment outcomes

Figures 1 and 2 plot the BDI and CORE-OM outcomes respectively for the 4 treatment conditions at the main assessment points. The plot for the BDI was not included in the original publication. Visually, these plots for the BDI and transformed CORE-OM are almost identical.

Table 1 shows the overall unadjusted means and standard deviations on the BDI and CORE for the full sample of 117 clients at Assessment 1, end of treatment and 3-month follow-up; also shown are prescreening scores on both measures, and pre-assessment to post-assessment (pre-post) effect sizes for the change from Assessment 1 to end of treatment, calculated as pre-assessment minus post-assessment means divided by the averaged pre- and post-assessment standard deviations. There were no gender effects.

Table 1: Overall means, standard deviations, and pre-post effect sizes for BDI-I and core OM

	Prescreening			**Assessment 1**			**End of Treatment**				**3 Month Follow-up**		
Measure	M	SD	n	M	SD	n	M	SD	n	ES	M	SD	n
BDI-I	24.5	6.3	110	21.4	6.8	117	9.5	7.7	113	1.77	10.2	8.7	115
CORE-OM clinical score	19.9	3.1	110	18.2	3.6	117	9.9	5.7	113	1.77	10.2	6.1	115

Treatment Effects

Tests of treatment modality main effect yielded similar results to those reported for the BDI (Shapiro et al, 1994), with a marginal effect in favour of CB: Madj 9.2, versus PI, Madj 11.6 ($F = 3.59$, df 1,97, $p = 0.06$).

Duration of Treatment

Although 16-session treatment held a numerical advantage over 8 sessions, this was not significant, just as reported for the BDI: 8 sessions, Madj 11.3 versus 16 sessions, Madj 9.4 ($F = 2.08$, df 1,103, $p = 0.152$).

Interactions between Treatment Duration and Severity of Depression

Table 2 shows adjusted means for the two comparison measures with the same significant result.

Table 2: Adjusted Means and Tests of Severity x Duration Interaction

	High Severity		Moderate Severity		Low Severity		P for simple effect of severity		Interaction effect		
Measure	8	16	8	16	8	16	8	16	F	dfs	p
BDI-I	19.8	9.3*	10.4	6.4	4.6	8.1	.002	.41	3.54	2,97	.03
CORE-OM clinical score	16.2	10.8*	11.4	7.9	7.1	10.2	.01	.63	3.51	2,103	.03

*** significant at the .05 level**

Discussion

The aim of the present paper was to test the proposition that rewiring an archived efficacy study, which originally used the BDI-I, with a new measure - the CORE-OM - via transformation formulae/tables would yield equivalent results to those originally reported (Shapiro et al., 1994, 1995). It is important to note that our selection of the SPP2 data set was driven by the availability of the dataset and in particular by our ability to replicate absolutely the detailed process of analyses that were originally carried out. Carrying out such a replication on an independent data set might have led to slight variation in procedures or analyses from those originally employed, thereby introducing a confounding effect. Overall, the results showed a virtually perfect replication of the four main findings and effects.

The importance of the present study lies in its implications for helping to bridge the gap between efficacy and effectiveness studies. Although there have been attempts to provide rules for transforming rating scale scores (e.g., Aiken, 1987), we are unaware of any other test in the psychological therapies literature in which an efficacy study has been rewired. In this respect, the results from the present study can only apply to transformations between the BDI-I and CORE-OM and vice versa. Transformations between the CORE-OM or BDI-I and other outcome measures would require the collection of new data and new transformation tables.

In terms of the present study, we highlight three implications. First, and crucially, it challenges the myth that an established outcome measure should always be used in efficacy studies because of the existing body of literature that has previously used that measure. Although the reasoning behind this assumption is sound, the results of the present study suggest that it is possible, using transformations based on large Ns, to adopt newer outcome measures without losing comparability with existing literatures. Practitioners carrying out research in routine settings might feel encouraged to select the CORE-OM as a research tool knowing that there is a mechanism for translating these scores so that comparisons can be made with studies using the BDI-I. Direct transformations can then be made between the BDI-I and BDI-II using the BDI-II manual (Beck et al., 1996).

Second, our findings highlight the potential for using rewired efficacy studies as benchmarks for current practice-based activity. Two examples can be drawn from the current findings whereby results could be extrapolated to different treatment packages of care that might be considered comparable to those considered in this study. First, in terms of the overall outcomes in SPP2, the clinical score was approximately 20 at screening and 18 at intake assessment and then fell to approximately 10 at end of therapy. Hence, the overall pre-post change is of the order of 8 points using the clinical scoring method (or 0.8 using the 0-4 scaling). This provides a global benchmark against which to compare both intake severity levels and outcomes in routine practice. Second, we can consider the extent to which therapies might be expect to differ based on the difference obtained here between PI and CB outcomes. Our findings yielded a difference on the clinical score in the

region of 2.5 between contrasting treatment approaches (or 0.25 if using the 0-4 scaling for the CORE-OM). Hence, the present study provides two initial benchmarks against which routine services can equate obtained effects in relation to (a) overall outcomes and (b) differences between types of therapies.

Third, at a technical level, the yield of a straight transformation of one measurement score to another would be unsurprising if there were a linear relationship between the two measures. However, while the BDI-I and CORE-OM are highly correlated in this and other samples, the empirical relationship is not linear. In addition, the two measures, while occupying the same conceptual space, differ in a fundamental way. The BDI-I is a specific measure of depression whereas the CORE-OM is a generic measure and its rationale was to tap the 'core' aspects of people's presenting problems (Barkham et al., 1998). The immediate implication is that the transformations yielded by the formulae and look-up tables reported in Leach et al. (in press) are sufficiently accurate and robust as to have widespread applicability in linking efficacy and effectiveness research.

Although we have developed a rewiring approach for transforming between BDI-I and CORE-OM scores, this is not the only means for obtaining a common metric. An alternative strategy would be to compare measures using standard scores (e.g., t or z scores) or effect sizes. However, although such procedures have existed for years, they are rarely used by practitioners. Part of the reason, perhaps, might be because such an approach effectively strips out the intrinsic or associated meaning captured by a particular score derived from a known measure. Most practitioners will, for example, have a tacit sense of the clinical gains implied in a BDI-I or BDI-II score moving from 32 at intake to 12 at discharge. By contrast, using a different case, the clinical meaning of reporting a pre-post effect size of, for example, 0.8 is less clear. The latter procedure masks the absolute levels at intake and discharge, thereby depriving the practitioner of valuable information. From a measurement perspective, standardised scores are useful for comparing between different measures when the precise relationship between those measures is not known. By contrast, the present procedures were possible because of precise transformations between the two measures drawn from the same sample of patients (Leach et al., in press). Not to use this information would entail losing a level of detail provided by the precision of comparisons gained by look-up tables that capture more of the fine detail of the relationship between the measures.

For both practitioners and researchers, the findings from the current study provide an empirical test of the precision of the transformations and evidence that using these transformations does not compromise the integrity of original findings. Recall also that the procedures used in this study could also be used to rewire a study originally using the CORE-OM and represent the results using the BDI-I. This may help to convince researchers either to adopt similar outcome measures as used by practitioners, or to rewire their analyses such that they present results in both original format (e.g. BDI) and transformed format (e.g. CORE-OM), thereby providing a key bridge between research and practice. Specifically, the findings reported here strengthen the potential relevance of an archival efficacy study to routine practice by translating its results into a metric that is widely used and hence readily interpreted by practitioners.

References

Aiken, L.R. 1987. Formulas for equating ratings on different scales. Educational & Psychological Measurement, 47, 51-54.

Barkham, M. & Mellor-Clark, J. 2000. Rigour and relevance: Practice-based evidence in the psychological therapies. In N. Rowland & S. Goss (ed.). Evidence-based counselling and psychological therapies: Research and applications (pp.127-144). London: Routledge.

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

Barkham, M., Evans, C., Margison, F., McGrath, G., Mellor-Clark, J., Milne, D. &

Connell, J. 1998. The rationale for developing and implementing core batteries in service settings and psychotherapy outcome research. Journal of Mental Health, 7, 35-47.

Barkham M, Gilbert, N., Connell, J., Marshall, C. & Twigg, E. 2005. Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. British Journal of Psychiatry, 186, 239-246.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K. & McGrath, G. 2001. Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. Journal of Consulting and Clinical Psychology, 69, 184-196.

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

Chwalisz, K. 2003. Evidence-based practice: A framework for twenty-first-century scientist-practitioner training. Counselling Psychologist, 31, 497-528.

Evans, C., Connell, J., Barkham, M., Margison, F., Mellor-Clark, J., McGrath, G. & Audin, K. 2002. Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. British Journal of Psychiatry, 180, 51-60.

Froyd, J.E., Lambert, M.J. & Froyd, J.D. 1996. A review of practices of psychotherapy outcome measurement. Journal of Mental Health, 5, 11-15.

Horowitz, L.M., Strupp, H.H., Lambert, M.J., Elkin, I. 1997. Overview and summary of the core battery conference. In H.H., Strupp, L.M. Horowitz, & M.J. Lambert (eds.), Measuring Patient Changes in Mood, Anxiety, and Personality Disorders : Toward a Core Battery. Washington, D.C.: APA.

Leach, C., Lucock, M., Barkham, M., Stiles, W.B., Noble, R. &, Iveson, S. (in press). Transforming between Beck Depression Inventory and CORE-OM scores in routine clinical practice. British Journal of Clinical Psychology.

Leach, C., Lucock, M., Iveson, S., & Noble, R. 2004. Evaluating psychological therapies services: a review of outcome measures and their utility. Mental Health & Learning Disabilities Research & Practice, 1, 53-66.

Shapiro, D.A., Barkham, M., Rees, A., Hardy, G.E., Reynolds, S. & Startup, M. 1994. Effects of treatment duration and severity of depression on the effectiveness of cognitive-behavioural and psychodynamic-interpersonal psychotherapy. Journal of Consulting and Clinical Psychology, 62, 522-534.

Shapiro, D.A., Rees, A., Barkham, M., Hardy, G.E., Reynolds, S., & Startup, M. 1995. Effects of treatment duration and severity of depression on the maintenance of gains following cognitive-behavioural and psychodynamic-interpersonal psychotherapy. Journal of Consulting and Clinical Psychology, 63, 378-387.

Waskow, I.E. 1975. Selection of a core battery. In I.E. Waskow & M.B. Parloff (Eds.), Psychotherapy change measures (DHEW Pub. No (ADM) 74-120). (pp.245-269). Washington, DC: U.S. Government Printing Office.

Acknowledgements

Authors affiliated to PTRC were funded by the Priorities & Needs R&D Levy via Leeds Community and Mental Health Teaching Trust.

Declaration of interest

Michael Barkham is a member of the CORE System Trust and received funding from the Mental Health Foundation to develop the CORE-OM.