Trial Design of Psychological Treatments in Chronic Pain: What Can We Tell Patients?

Stephen Morley

Leeds Institute of Health Sciences, University of Leeds, Leeds, United Kingdom

It is 40 years since the publication of replicated single-case studies reporting the effectiveness of applied behavior analysis to the treatment of pain and almost 25 years since the publication of the first, albeit small, randomized controlled trials (RCTs) of cognitive behavioral therapy (CBT) for chronic pain. Today there is a substantive body of RCTs—upward of 50 trials—examining the efficacy of psychological treatments, predominantly CBT, for chronic pain. Several systematic reviews of this literature have been produced, and the broad conclusion reached by these studies is that psychological treatment “works” (Morley et al. 1999; Guzmán et al. 2001; Astin et al. 2002). Considered as a whole, the trials are markedly heterogeneous over a range of domains (Morley and Williams 2002), with variations in the package of CBT offered as treatment; the length and number of sessions of treatment; the context in which treatment is delivered, ranging from highly specialized in-patient care (Williams et al. 1996) through community settings (Von Korff et al. 2005) to Internet delivery (Buhrman et al. 2004); the diagnostic group recruited to the study; and the outcome measures used. One might be encouraged by this heterogeneity in that it could be construed to provide support for a generalized conclusion that as a class of treatment, CBT is effective for chronic pain—especially as many of the studies were conducted in clinically relevant conditions (c.f. Shadish et al. 1997).

From the perspective of a practicing clinician, what should I tell the patients that I see? I doubt if telling them that on average they might expect to be at the 69th percentile of the average patient in a waiting-list control group would be perceived to answer their query, and a detailed account of the evidence base would be similarly inappropriate. Patients want to know what objective benefits
they might expect to experience. At present it is difficult to say “If you have this treatment” (not if you have this therapist) “then you will, in all probability, gain these specified benefits.”

In this chapter I focus on three areas in the current state of trials and highlight where I believe we, the clinical research community, should be concerned: measurement and the meaningfulness of change, control groups and what we can conclude from them, and the quality of trials with particular reference to treatment. The points that are raised are not unique to psychological treatments for chronic pain or even for psychological treatments per se. Contemporary psychological treatments for chronic pain have their roots in more general approaches to ameliorating distressing experience and behavior. It is therefore not surprising that many of the issues involved in evaluating the effectiveness of treatments for chronic pain reflect similar issues in the field of psychotherapy research. Many of the issues remain the subject of lively discussion, and the challenges that face the accumulation and interpretation of the evidence base in pain are shared.

MEASUREMENT AND MEANINGFULNESS OF CHANGE

The epidemiologist Geoffrey Rose noted that “disease truly forms a continuum of severity but its management requires a system of unambiguous labels,” jesting that “doctors count to 2” (Rose 1992); for an example in pain see Moore et al. (1997). In contrast, the evaluations of psychological interventions have generally acknowledged the continuity of psychological variables and have pursued a strategy that relies on inferential statistical testing of between-group differences based on continuous measures. This situation is endemic in RCTs of psychological treatments for pain (Turk 2000). Almost without exception, results are reported with reference to P values for differences between means rather than differences in categories that relate to criterion events. I suggest there are two interrelated issues to be grappled with: (1) establishing other criteria for evaluating data and (2) enhancing the meaning of the measures used.

The top half of Fig. 1 illustrates some key aspects of the prevailing approach to measuring psychological variables. Many variables cannot be observed directly and we assess them by selecting meaningful items that validly represent the construct and by using defined psychometric methods to establish a scale and response format. While such scales are internally coherent and reliable we often do not know how well they map onto the underlying dimension. In Fig. 1, scales X and Y—e.g., measures of catastrophizing, coping, or depression—may cover different ranges of the relevant variable, and there is no way of knowing that zero on either scale corresponds either to the other scale or to
the underlying scale. The units on each scale will also vary as function of the number of items and response formats used in the scales. It is possible to map scores from one scale to the other using regression. The assumption of equal intervals within a scale is often made, but it is difficult to check, and scale Z probably offers a truer picture of measurement. Exceptions to these general statements may be found in psychophysical scaling when subjective experience can be calibrated against an external stimulus—a well-tried procedure in experimental pain (Gracely 2005).

These issues are widely acknowledged in psychology, and despite the arbitrary nature of this type of measurement (Blanton and Jaccard 2006), the measures function well in many settings where theories and models are tested. A problem arises when the measures are used in applied settings (Kaplan 1990; Sechrest et al. 1996; Kazdin 2006), where the calibration between scale scores and specific, important behavioral referents has generally not been addressed. In passing it is worth acknowledging that studies conducted within the framework of applied behavior analysis generally obviate the problems raised by the types of measures under discussion because they focus directly on the criterion behavior of interest (c.f. Fordyce 1976). A similar position has arisen with the highly specified treatment derived from the fear-avoidance model (Vlaeyen et al. 2001; de Jong et al. 2005).

Fig. 1. Upper half: Schematic representation of measurement (scaling) issues. Lower half: Schematic representation of clinically significant change criteria. CSC = clinically significant change. See text for detailed commentary.
An increasingly popular approach to defining meaningful change uses the psychometric properties of scales and normative data to establish criteria for a reliable change index (RCI) and clinically significant change (CSC). The method was popularized by Jacobson (Jacobson and Truax 1991; Jacobson et al. 1999) and has received considerable attention (Journal of Consulting and Clinical Psychology 1999; 63[3], special issue) and technical debate (e.g., Atkins et al. 2005). Information on the reliability of the scale and the variability of the sample is used to determine the magnitude of change that can be regarded as reliable, i.e., not due to variation caused by the unreliability of the scale. Individuals can thus be classified as reliably improved or not. Knowledge of the distributions (assumed to be normal) of the test scores is used to define cut scores to determine whether observed changes are “clinically significant.” Jacobson proposed three possible criteria, which are illustrated in the bottom half of Fig. 1. Criterion (a) is whether subsequent to therapy the individual’s score falls outside the range of the dysfunctional sample, defined as within the 2.5% of the relevant tail of the distribution of the dysfunctional sample. Criterion (b) is whether the individual’s score falls within the range of the functional sample, defined as the central 95% of the distribution of the functional sample. Criterion (c) is whether, after treatment, the individual’s score is closer to the mean of the functional sample than the dysfunctional sample.

Figs. 2 and 3 show a graphic representation of RCI/SCS methodology applied to a sample of more than 800 patients receiving treatment in a 4-week CBT-oriented inpatient program (INPUT, London; S. Morley et al., unpublished

![Tramline display for RCI and CSC](image)

**Fig. 2.** A tramline display for the reliable change index (RCI) and clinically significant change. See text for detailed commentary.
The data were collected 1 month after the end of treatment and therefore show the immediate post-treatment effect of the program. By conventional statistical criteria, the program would be considered outstandingly successful: for all pre- to poststatistical comparisons, $P < 0.001$. Applying the RCI/SCS method provides a more nuanced view of the effectiveness of the program. Fig. 2 represents both RCI and CSC in a “tramline” display in which the pretreatment scores are plotted against post-treatment scores. In the case of perfect reliability of measurement and no change the data would be perfectly linear, as shown by the line that diagonally bisects the plot. The shaded symmetrical area around this line represents the variability of scores attributable to the unreliability of the measure (in the present case, the 99% CI has been set at ± 5 scale points)—the RCI. Superimposed on this display is the cut score for a predetermined CSC (set at 15). The areas formed by the RCI and the cut scores define several possible categories of outcome. At the right-hand side of the graph are shown three individuals who scored 25 before treatment. Post-treatment there are three outcomes: reliable change but not clinically significant improvement, reliable change and clinically significant improvement, and clinically significant improvement.
reliable deterioration. The left-hand side shows changes in three people whose pretreatment scores are below the cut-off. Two of these (reliable improvement and reliable deterioration) may be relatively trivial because the patient still remains below the cut-off score, but the third (reliable change and above the cut-off score) represents a definite clinical deterioration. In Fig. 3, a series of stacked bar charts displays the percentage of patients achieving each of the outcomes for the nine outcome measures used. The data represent those patients who were above the cut scores for each scale at pretreatment, i.e., those on the right-hand side of Fig. 2. For many variables more than half the sample did not experience a reliable change; a smaller number made reliable changes that did not meet criteria for clinically significant change; there is variation across the measures in the proportion of individuals making clinically significant improvement; and finally, the data suggest that for every measure there is a small proportion of people who reliably deteriorated over the period of treatment. Variation across the measures in the numbers of individuals falling into each category is not only a possible function of treatment but is also related to variations in the reliability of the measures, the variance of the sample on the measures, and the criterion (a, b, or c) applied. Selection of the relevant criterion depends on the availability of data from relevant samples; the separation or overlap of the distributions will influence the interpretation placed on the outcome. Notwithstanding these caveats, RCI/CSC provides a rational methodology for dividing the continuum that defines clinical significance by the intrinsic statistical properties of the scale. Additional meaning must be sought by cross-referencing the norms to psychological and behavioral referents of interest (Kazdin 1999).

Clinical significance has several other meanings, including absence of disease—the elimination of the relevant pathogens; risk reduction of other harm—return of a biological marker to a point that signifies the reduction in the risk of acquiring additional life-threatening disease; and reduction in the probability of relapse. The relevance of each criterion is contextually dependent on the condition and treatment. The reader is invited to consider leukemia, AIDS/HIV, or major depressive disorder to generate appropriate clinical significance criteria. It is may also be necessary to know the details of the individual’s clinical history and previous treatment before interpreting a score. For example, the meaningfulness of a low BDI score in patients with prior episodes of depression as it relates to relapse depends on the number of previous episodes and the treatment received (Teasdale et al. 2000).

Pain researchers are beginning to address the issue of meaningful change, at least with respect to the experience of pain intensity, by devising patient defined criteria for meaningful change (Farrar 2000; Farrar et al. 2001, 2003; Cepeda et al. 2003a, b; Cucchiaro et al. 2006; Hanley et al. 2006). The main method is
to correlate change scores (pre- to post-treatment) against a categorical scale assessing the patient’s judgment about whether the change in pain is noticeable and whether the pain has decreased to a meaningful degree. The apparent simplicity of this method belies the complex cognitive judgment required. Memory, the requirement of patients to generate a criterion of “meaningfulness,” and the variation in scales used, all contribute to performance on the task. Unsurprisingly, there is often considerable variance in the data, and the magnitude of the meaningful change is a function of the magnitude of initial pain intensity. On the basis of present evidence it seems that patients require a reduction of pain intensity somewhere between 30% to 50% for it to be regarded as meaningful. This approach to defining meaningful change is to be welcomed, and it should be extended to measures considered important in the treatment of pain (Turk et al. 2003). Other “stakeholders,” such as families and employers, might also be consulted to determine the relevance of outcome measures, which have thus far been selected by the clinical research community (Morley and Williams 2002). Research on human judgment suggests that “quantitative subjective assessments are almost always biased, sometimes completely misleading” (Poulton 1977), and future research into stakeholders’ judgments of clinical or meaningful change should be informed by the extensive psychological theory and methodology available (Poulton 1989).

THE ISSUE OF CONTROL GROUPS

shows the frequency with which different types of control groups were used in 31 RCTs of CBT analyzed by Yates et al. (2005). The total of 36 control groups is attributable to the fact that several studies used more than one control group. Table I also indicates the function of the control groups in ruling out plausible rival hypotheses. The addition of any well-matched randomized control group that is measured at the same points in time as a treatment group effectively deals with two major rival hypotheses that might account for change in the treated group: (1) the passage of time and the associated spontaneous remission of the problem, and (2) measurement artefacts (testing, instrumentation, and regression to the mean) (Cook and Campbell 1979). The use of waiting-list controls in chronic pain may be equivalent to an active “treatment as usual” (TAU) control because patients continue to receive care, self-medicate, or self-manage their pain; data are not always available to assess these possibilities. Authors appear to use a TAU arm where patients are either explicitly told to continue a preexisting treatment or where they are allocated to a standard (often medical) treatment. In the latter case it is possible that the allocation to TAU creates renewed expectations of improvement and that this condition
effectively controls for expectancy of treatment gain. It is unclear how well treatments involving education and support and attention control also control for equivalent expectations.

The control groups used do enable us to conclude that psychological treatment does produce genuine change, i.e., the difference is not an artefact of measurement, but at present it is difficult to be sure that observed improvements are attributable to the specific mechanisms suggested by the treatment rationales—the specific ingredients model. This position is compounded by the relative paucity of studies of putative change processes in psychological treatments for pain. Most published analyses report correlations between changes in process and outcome variables within single cohorts, and, while indicative, such analyses do not enable one to make strong causal conclusions (Morley 2004). Analysis should examine lagged changes between process and outcome variables controlling for autocorrelations, and more crucially demonstrate an interaction between the specified process and the delivery of treatment, i.e., the critical process occurs in the treatment group but not the control (Kraemer et al. 2002). Turner and her colleagues (2007) have reported such an analysis of a recent trial (Turner et al. 2006) comparing CBT with an education/attention control condition. This pioneering study does report a critical association between changes in theoretically relevant cognitive variables and outcome variables, and it tests for the critical interaction between this association and treatment.

While Turner et al.’s (2007) analysis intimates support for the therapeutic rationale, modifications in the design and analysis of trials will be necessary to test fully the specific ingredients model of treatment. The alternative hypothesis

Table I
The frequency of control groups used in 31 RCTs for psychological interventions and the extent to which specific threats to validity are controlled

<table>
<thead>
<tr>
<th>Control Group</th>
<th>Frequency</th>
<th>Threat to Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waiting list</td>
<td>16</td>
<td>+</td>
</tr>
<tr>
<td>Treatment as usual</td>
<td>9</td>
<td>+</td>
</tr>
<tr>
<td>Education</td>
<td>4</td>
<td>+</td>
</tr>
<tr>
<td>Attention control</td>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>Symptom monitoring</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>Physiotherapy</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>Exercise</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>Bibliotherapy</td>
<td>1</td>
<td>+</td>
</tr>
</tbody>
</table>

Note: Data from Yates et al. (2005). A plus sign (+) indicates that the threat to validity is controlled for by the named control group; a question mark (?) indicates that the threat may be adequately controlled in some circumstances.
is that change is attributable to general contextual factors associated with the delivery of therapy. The common factors hypothesis has a substantial history (Frank and Frank 1991), and a compelling contemporary account of the argument is given by Wampold (2001). Wampold’s case is made largely on the treatment for depression, where there are extensive high-quality primary data comparing bona fide treatments and relevant meta-analyses. Of the many points made by Wampold, two have particular relevance to the pain literature because they appear not yet to have been fully addressed.

The first concerns the design of placebo treatments to control for common factors present in treatments. Wampold and his colleagues (Baskin et al. 2003) observed a significant discrepancy in the design of control treatments in that many of them are structurally nonequivalent. That is, they are not matched to the treatment arm with respect to “number and duration of sessions, training of therapist, format of therapy” (p. 973). Using meta-analysis these investigators showed that comparisons between active treatments and placebo treatments that were not structurally equivalent produced larger effects ($M = 0.465$, $95\%$ CI = 0.309, 0.621) than did comparisons between active treatments and structurally equivalent placebo treatments. More importantly, the latter comparison (treatment vs. structurally equivalent controls) generated a negligible effect ($M = 0.149$, $95\%$ CI = 0.005, 0.292) that was significantly smaller ($P = 0.003$).

The authors concluded that active treatments were not demonstrably superior to well-designed placebo conditions. Indeed, the very small effect size for the second comparison indicates that the specific effects of treatment also may be negligible, a point persuasively argued by Wampold (2001). We have yet to test this hypothesis for the treatment of chronic pain, and the control groups used in studies so far—see Table I—suggest that the present state of affairs means that we cannot exclude a common factor explanation for the effect of treatments.

The second factor neglected in the current literature is the relative influence of therapists rather than therapy. There is evidence that particular therapists produce better consistently better outcomes than others. In a recent reanalysis of the well-known NIMH Depression Collaborative Research Program (Kim et al. 2006), estimates were made of the proportion of variability in outcomes resulting from therapists. These authors used multilevel analyses to determine therapist effects that were nested within treatment arms (for a full discussion of this complex issue readers are referred to a series of papers in *Psychotherapy Research* 2006; 16(2):143–187). Kim and colleagues concluded that about 8% of the variance in outcomes was attributable to therapists while 0% was due to the particular therapy delivered. The latter finding replicates earlier analyses of the trial, where no differences between the treatments were observed. Moreover, these effects were detected in an RCT in which therapists used a manual and received appropriate supervision and training—procedures designed to assure
treatment (therapy) fidelity and to reduce between-therapist variance. Similar therapist effects were detected in a large clinical database of 6,000 patients treated by 580 therapists (Wampold and Brown 2005), suggesting that therapist effects are prevalent across both RCTs and clinical practice. At present we do not know if these findings apply to the treatment of chronic pain, but it would be surprising if this were not the case. The analysis of such effects requires careful consideration, and the issue is further compounding by the fact that psychological treatment for chronic pain is delivered in group format. Trial analysts have rarely considered this feature (Baldwin et al. 2005). Appropriate analyses would result in a reduced estimate of the size of effects attributable to the treatment component and direct our attention to neglected therapist (group) effects.

QUALITY OF TRIALS: ASSESSING TREATMENT

The importance of trial quality as a source of bias affecting outcomes is well recognized, and a number of scales have been devised to assess various aspects of trial quality. The majority of them focus on key aspects of design that define modern RCTs: randomization; blinding of participants, therapists, and assessors; statistical power; and integrity of statistical analysis. It is accepted methodology for meta-analysts to include an assessment of trial quality. The resulting quality score either is used to exclude trials from the meta-analysis, or can be entered as a covariate so that the influence of quality (and its various components) on observed effect sizes can be examined. The application of many of the standard quality scales to trials of psychological treatments is problematic because some of the central tenets of good trial design are difficult, if not impossible, to meet. For example, neither therapists nor patients can be “blinded” to treatment, and blinding of assessors is also problematic. Thus, application of standard quality scales to psychological treatment trials can “bias” conclusions about their quality (Guzmán et al. 2001; Price et al. 2001). In an earlier meta-analysis of psychological treatments, we (Morley et al. 1999) decided not to use a quality scale because of this potential problem. This limitation, and the relatively small number of trials available, also meant that we were unable to systematically analyze the influence of quality on outcome.

Correction for this problem might be made by incorporating additional methodological elements into the assessment of trial quality. For example, one purpose of blinding patients and therapists to treatment in conventional trials is to control differential treatment expectations and the consequent demand characteristics that may be generated. Within the field of psychological treatment there is a long tradition of directly assessing treatment credibility and expectation and statistically controlling for differences in the final analysis (Borkovec
and Nau 1972; Kazdin and Wilcoxon 1976) and the inclusion in a quality scale of an item assessing this strategy would mitigate the problem of standard quality scales. The rationale for this is supported by emerging evidence that pretreatment expectations do make a modest contribution to the variance of outcome scores (Goossens et al. 2005; S. Morley et al., unpublished manuscript).

Recently Yates et al. (2005) described the development of a scale for assessing the quality of psychological trials for chronic pain. A Delphi method was used to consult a group of 15 international experts (published trialists). After three consultation rounds, reasonable consensus was reached on 84 statements representing desirable quality. There was a certain redundancy in the statements that enabled an expert panel to pool them to form a 27-item scale. Twenty-one of the items represented aspects of good practice in design and methodology, and six items examined the implementation of treatment (treatment quality). Table II provides a summary of the reliability of the scale items for 31 trials and three raters. The reliability for the total scale assessed by intraclass correlation (ICC) was 0.91 (95% CI = 0.76, 0.96) for the full scale, 0.91 (95% CI = 0.76, 0.96) for the treatment subscale, and 0.85 (95% CI = 0.70, 0.93) for the design and methods subscale. To test the general utility of the scale, five novice raters achieved more modest reliability: the median ICCs for 5 pairs of “novice” raters were 0.81 for the total score, 0.57 for the treatment subscale, and 0.76 for the design and methods subscale. A provisional attempt to establish the validity of the scale was made by regressing the quality score onto the year of publication for each trial. The expectation that trial quality, or the reporting of trials, should improve over time was supported, and the ability of the scale to discriminate between predefined “poor,” “average,” and “excellent” trials was observed. The scale therefore seems to provide a valid tool for investigating the influence of quality factors in psychological treatments for pain.

Of interest is the frequency with which individual criteria were met across the trials. This percentage is shown in the last two columns of Table II. Trials were generally good at reporting the content of treatment, sample details and inclusion/exclusion criteria, details of outcomes, and analyses, but notably deficient in some technical areas of analysis, such as intention-to-treat analysis and power calculations. There is marked variation, particularly in the measurement of treatment quality. The relevance of this variation is that psychological treatment is essentially “manufactured” on each and every occasion that it is delivered. By way of contrast, the quality of pharmacological treatments is controlled in the manufacturing process, and provided the health provider ensures that the medicine is bona fide, the quality of treatment is guaranteed. (This example ignores issues of adherence and that the way in which doctors give the treatment might make a difference, but the specific ingredient of therapy is guaranteed.) Psychological treatments are different in that they are
complex, multi-componential, interpersonal strategies that have to be adjusted to variations between individuals. In order to interpret trials we need to know what was delivered, by whom, whether the quality of treatment was acceptable,
and whether patients adhered sufficiently to the treatment. Additionally, for the treatment to be applicable in clinical settings, explicit studies of its dissemination are required. Borrelli et al. (2005) have developed a more detailed tool to assess treatment fidelity that elaborates the brief scale of Yates et al. (2005). Borrelli et al.’s scale separates treatment fidelity into five components: treatment design, training of therapists, delivery of treatment, receipt of treatment, and enactment of treatment skills. The latter two items are explicit attempts to assess patients’ “uptake” of the treatment and their active use of treatment strategies. In an examination of 342 trials, Borrelli et al. (2005) documented the frequency with which various criteria were met and the reporting of treatment fidelity criteria over a period of 10 years. From the perspective of pain research, it is cold comfort that trials in the general field of behavioral health also struggled to guarantee treatment quality and that there was no discernible improvement in reporting over time.

CONCLUDING COMMENTS

Designing high-quality trials of psychological interventions is undoubtedly challenging. While there is evidence that treatment is statistically more effective than no treatment, the way in which the majority of trials have been designed make it difficult for us to tell a patient exactly what benefits will accrue. This difficulty may exist partly because chronic pain is a fuzzy multidimensional construct (as indicated by the variety of measures). Undoubtedly, experienced individual clinicians can, from their knowledge base, give a reasonable account of what a patient seeking treatment might expect to gain, but we still have to capture this expectation in our measurement and evaluation strategies. More intriguingly, the present designs do not enable us to say why the treatment works. While the pursuit of the specific ingredients model, e.g., change being mediated by particular cognitive or behavioral processes, will be the preferred modus operandi, we should be alert to the general influences of context (therapist effects). We may yet have to advise patients to find a good therapist (or therapeutic team) rather than a specific therapy (Wampold 2001).

ACKNOWLEDGMENTS

Many thanks to my colleague Dr David Owens (University of Leeds) for discussions on clinical significance, and for introducing me to the work of Geoffrey Rose; to Dr Jeff Lackner (SUNY-Buffalo) for comments on an earlier version; to Dr Fiona Smith for discussion on issues of meaningful change in pain; to Dr Sumerra Hussain for analysis of the RCI/CSC data; also, to my
partners in crime over many years, Amanda Williams and Chris Eccleston—I remember a particular “kidnapping” in Verona—“thank you” from a willing victim.

REFERENCES


Cepeda MS, Africano JM, Polo R, Alcala R, Carr DB. Agreement between percentage pain reductions calculated from numeric rating scores of pain intensity and those reported by patients with acute or cancer pain. *Pain* 2003a; 106:439–442.


**Correspondence to:** Stephen Morley, PhD, Leeds Institute of Health Science, University of Leeds, Charles Thackrah Building, 101 Clarendon Road, Leeds LS2 9LJ, W. Yorkshire, United Kingdom. Tel: +44-113-343-2733; fax: +44-113-243-6997; email: s.j.morley@leeds.ac.uk.