Journal of Traumatic Stress, Vol. 20, No. 5, October 2007, pp. 779–792 (© 2007)

The Rocks and Hard Places in Psychotherapy Outcome Research

Paula P. Schnurr
National Center for Posttraumatic Stress Disorder, White River Junction, VT, and Department of Psychiatry, Dartmouth Medical School, Lebanon, NH

The field of traumatic stress has experienced significant growth in the number of psychotherapy treatment studies conducted, especially in the past 10 years. In this article, the author aims to help both the readers of this work and psychotherapy researchers by reviewing key concepts in trial design and discussing issues unique to psychotherapy (and other nonpharmacological interventions). The discussion is broad and nontechnical, particularly with respect to statistical issues. The focus is on between-group randomized designs, and on outcome rather than process. Examples are drawn from posttraumatic stress disorder treatment studies to enhance relevance to the field of traumatic stress.

In an important article in the field of psychotherapy outcome research, Borkovec (1993) began by stating, "the fundamental goal of any between-group experimental design and its associated methodology is to hold all factors consistent other than the one variable about which cause-and-effect conclusions are to be drawn" (p. 249). In short, the issue is control.

All treatment research—indeed all research—requires strategies to control factors that influence the validity of inferences drawn from the findings. However, psychotherapy research presents greater challenges to achieving control relative to research involving medication treatment. Psychotherapy research has difficulties in controlling patient and provider knowledge of the treatment the patient receives, the effect of patient–provider interactions on outcomes, provider adherence to treatment protocols, and standardization of treatments. Blinding patients and providers to which treatment the patient is receiving is virtually impossible in a psychotherapy study, unlike the blinding that is possible when identically appearing drugs and placebo pills are administered in a placebo-controlled design. In fact, there is no true placebo in a psychotherapy study. Patients assigned to a wait-list receive no treatment as part of the study protocol, and control treatments have active elements. The psychotherapeutic intervention itself requires repeated interaction between patients and providers. Although medication visits in a drug study could influence patient outcomes, the drug, and not the interaction between the provider and patient when the drug is administered, is the primary treatment. A psychotherapist's skill can influence the delivery of the intervention. A provider in a medication study might fail to follow the study's dosing protocol, but the complexity of following the protocol is unlikely to be as great as the complexity of following a multisession psychotherapy manual. Monitoring provider adherence is relatively more important for a psychotherapeutic intervention. The quality control built into the process of manufacturing a drug to standardize dose and

Correspondence concerning this article should be addressed to: Paula P. Schnurr, National Center for PTSD (116D), VA Medical Center, White River Junction, VT 05009. E-mail: Paula.P. Schnurr@Dartmouth.edu.

© 2007 International Society for Traumatic Stress Studies. Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/jts.20292

appearance is not possible in psychotherapy. Each session differs in a way that each pill does (or at least should) not.

The comparison between psychotherapy and drug studies is not meant to imply that the latter are easy, or that the issues in psychotherapy are unique. As indicated above, the difference is more quantitative than qualitative. Moreover, the challenges in psychotherapy research apply broadly to all nonpharmacological interventions, such as surgery and physical therapy too. In fact, the similarity of surgery research is indicated by the novelty of a trial that employed sham surgery to determine the benefit of arthroscopic treatment for osteoarthritis of the knee (Moseley et al., 2002). It is difficult for surgery researchers and psychotherapy researchers alike to use placebo controls.

The field of traumatic stress has experienced significant growth in the number of psychotherapy treatment studies conducted, especially in the past 10 years. This article aims to help both readers of this work and psychotherapy researchers by reviewing key concepts in trial design and discussing issues unique to psychotherapy (and other nonpharmacological interventions). The discussion is broad and nontechnical, particularly with respect to statistical issues. Like Borkovec's (1993) article, the focus is on between-group randomized clinical trials (RCTs), and on outcome rather than process. Examples are drawn from a large multisite trial of treatment for posttraumatic stress disorder (PTSD) in female veterans and active-duty personnel, Department of Veterans Affairs Cooperative Study #494 (VA CSP #494; Schnurr et al., 2005, 2007) and other PTSD treatment studies to enhance relevance to the field of traumatic stress.

KEY CONCEPTS IN TRIAL DESIGN

In 1995, Seligman's landmark article in the *American Psychologist* on the benefits of psychotherapy drew attention to the concepts of efficacy (whether a treatment works) and effectiveness (whether the treatment works in practice). Psychotherapy efficacy studies include features intended to control threats to internal validity (i.e., threats to inferences about cause and effect) and construct validity (i.e., threats to the operationalization of the independent and dependent variables). These features include random assignment; rigorous control for the nonspecific benefits of treatment; manualized treatment and monitoring therapists' fidelity to the manual, usually by videotaping sessions; a fixed number of sessions; well-defined outcomes; blind raters; the exclusion of patients with multiple comorbidities; and fixed assessment intervals. Seligman argued that such research failed to capture psychotherapy as done in practice and emphasized the need for effectiveness studies, which include few or none of the characteristics of efficacy designs. For example, an effectiveness study might not use randomization and or manuals and might use patient-specific assessment intervals (such as when the patient and therapist feel treatment goals have been met).

It is tempting to believe that effectiveness studies are uninterpretable, or that efficacy studies have limited practical utility. In fact, there is no absolute advantage of one type of study over the other. Efficacy studies have greater internal and construct validity because of design features that create homogeneous treatment groups, standardized treatments, careful measurement of treatments and outcomes, and fewer opportunities for confounding effects of variables other than the treatments. However, effectiveness studies have greater external validity (i.e., generalizability) because they more closely resemble clinical practice. Therefore, the key issue when judging the merits of any particular study is how much we know about the treatment. Rounsaville, Carroll, and Onken (2001) have proposed a three-stage framework for evaluating treatments that illustrates this "right tool for the job" idea. The initial stage is used to establish feasibility and conduct pilot studies. In the second stage, RCTs establish the treatment's efficacy. Effectiveness studies are used in the third stage to establish the actual benefit in clinical settings. With this staged approach, the relative advantages of each design incrementally enhance the spectrum of knowledge about the treatment of a given condition.

Practical Trials

Another key concept in trial design is the practical clinical trial, developed to answer questions faced by those who

need to make decisions about healthcare, including patients and their clinicians as well as payers, administrators, and policymakers (Tunis, Stryer, & Clancy, 2003). Such trials are distinguished from explanatory trials, which aim "to better understand how and why an intervention works" (p. 1626). Practical trials combine elements of efficacy and effectiveness designs. They are RCTs with clinically relevant comparison groups, diverse patient populations, heterogeneous practice settings, and measurement across a range of outcomes.

For example, CSP #494 (Schnurr et al., 2005, 2007) was designed to produce information that could be used to inform decisions about cognitive–behavioral therapy for PTSD in VA and Department of Defense settings. At 12 sites across the United States, 284 female veterans and active-duty personnel were randomly assigned to receive either Prolonged Exposure (Foa & Rothbaum, 1998) or a Present-Centered Therapy that offered a credible therapeutic alternative to control for the nonspecific benefits of treatment. The sites included nine VA hospitals, two VA community Vet Centers, and one military hospital. The 52 therapists ranged in expertise with cognitive–behavioral techniques; most were not experts.

Inclusion and exclusion criteria were designed to maximize patient heterogeneity. Most types of Axis I and all types of Axis II comorbidity were allowed. Exclusion criteria were limited to factors that would affect patient safety (e.g., serious suicidality, living in a violent relationship) or the feasibility of delivering Prolonged Exposure therapy (e.g., significant cognitive impairment, no clear traumatic memory). Patients were allowed to use medications, attend self-help groups, and receive therapy for conditions other than PTSD. Measures included PTSD symptoms, depression, anxiety, substance abuse, functioning, quality of life, satisfaction, and utilization.

Like the distinction between efficacy and effectiveness, the distinction between practical trials and those with more limited utility for decision makers is a matter of degree. Many RCTs include features of practical trials. Their relevance for making healthcare decisions is enhanced to the extent that they do.

Intention-to-Treat Analysis

A lot can go wrong in a clinical trial. Patients may never show up for treatment, drop out before treatment is completed, and obtain the treatment to which they are not assigned from a source outside of the study. They may fail to complete assessments too. Patients are not the only problem. Therapists may fail to deliver the treatment as intended or use prohibited techniques from another condition. Such protocol deviations present a challenge for data analysis, causing researchers to analyze the data from patients who receive treatment per protocol, such as treatment completers or patients who receive a minimum amount of treatment (e.g., 24 of 30 group sessions; Schnurr et al., 2003). This strategy, however, threatens internal validity because it tampers with randomization. Yet is it fair to include data from treatment dropouts or patients who received suboptimal therapy, knowing that this will attenuate the treatment effect (if the treatment is effective)? The answer is yes; in fact, this yields the most unbiased effect (e.g., Lachin, 2000) because the analysis is grounded in the randomization.

This strategy, in which participants' data are analyzed according to their group assignment regardless of whether they received treatment as intended, is known as *intention-to-treat*, or ITT, analysis. Once randomized, each patient is included in the analysis according to group assignment. Dropouts cannot be replaced and even additional treatment stipulated as an exclusion criterion cannot be disallowed (although study treatment may be discontinued if appropriate).

The ITT principle is sometimes misunderstood. Studies in which data from some postrandomization dropouts are excluded may be described as using ITT analysis (e.g., Foa et al., 2005). In fact, several colleagues have reacted with incredulity upon hearing that ITT analysis actually requires including patients who do not receive any treatment. This seems to contradict the goal of determining if a treatment works: How can a treatment's effect be gauged if data from untreated patients must be included? That is the wrong question. The better question is, "What do we need

to know?" Intention-to-treat analysis provides the best estimate of the treatment effect in the population of patients to whom a study's results can be generalized. Thus, in an ITT analysis, the question is, "Is it better to adopt a policy of Treatment A if possible, with deviations as necessary, or a policy of Treatment B if possible, with deviations as necessary, for patients who seem to have this disease?" (Peto et al., 1977, p. 29). It estimates the likely benefit when the treatment is delivered in practice.

A word of caution for readers who may assume that ITT analysis is for effectiveness studies or practical trials and that completer analysis is acceptable for efficacy studies. Although per-protocol analysis is more relevant for efficacy studies, it can undo the benefits of randomization. One can never be certain that there is not some unmeasured factor that is wholly or at least partially responsible for differences observed between a treatment and a comparison condition. Per-protocol analysis should also be treated as what it is—an unplanned subgroup analysis—and interpreted carefully. Making correct inferences can be facilitated by using statistical strategies that have been developed to minimize bias in estimating the effects of treatment received per protocol (see Goetghebeur & Loeys, 2002, for a brief summary).

One final note regarding compliance: Noncompliance with treatment is not the same as noncompliance with measurement. Every attempt should be made to measure all participants, regardless of whether they discontinue treatment. Measuring treatment dropouts greatly facilitates data analysis because it removes the confound between dropout and missing data.

SPECIAL CONSIDERATIONS IN PSYCHOTHERAPY TRIALS

Comparison Conditions

Types of comparison conditions. As mentioned above, it is not possible (or perhaps, it is rarely possible) to use a placebo control in a psychotherapy trial. Borkovec (1993) describes six basic designs, three of which involve systematic manipulation of some elements of the treatments being compared; these three designs are described collectively below.

The simplest design employs an untreated comparison group that is assessed at the same intervals as the treated group(s) (e.g., Cloitre, Koenan, Cohen, & Han, 2002; Foa et al., 1999; Monson et al., 2006). This design is often described as a *wait-list design* because it is ethically appropriate to offer these patients treatment after the assessment period has ended. As Borkovec notes, a wait-list design controls for most threats to internal validity as delineated by Cook and Campbell (1979), including history, maturation, testing, instrumentation, statistical regression, selection, mortality (attrition), as well as interactions of these threats with selection. However, these designs permit no inferences about mechanism and whether there is anything specifically beneficial about the target treatment(s).

A nonspecific comparison design employs a comparison group that receives therapy that does not include what are considered to be the mechanisms of action in the target therapy (e.g., Schnurr et al., 2003, 2007); an alternative is a comparison group that receives treatment as usual in the setting where the research is conducted (e.g., McFall et al., 2005). These designs are used to control for the nonspecific benefits of treatment and permit inferences about the additional benefit of receiving the target treatment(s). For example, the Present-Centered Therapy in CSP #494 excluded exposure or other cognitive–behavioral techniques, as well as trauma focus. The greater improvement among patients who received Prolonged Exposure therapy thus could be specifically attributed to the type of treatment, and not merely to the fact that patients received treatment. However, Prolonged Exposure (the treatment protocol) included not just Prolonged Exposure (the technique), but also education, breathing retraining, listening to tapes of Prolonged Exposure therapy sessions as homework, in vivo exposure, and discussion of thoughts and feelings related to the exposure exercises. Present-Centered Therapy also included education and homework, although the type of homework differed from that used in Prolonged Exposure therapy. With so many differences between treatments, it is not possible to determine which of the elements in

Prolonged Exposure therapy were responsible for the additional improvement it produced.

Another class of designs is used to address this question. These designs involve systematic manipulation of the elements thought to be the mechanisms of a treatment's action. In a dismantling (component control) design, these elements include some to all of the elements to isolate the essential element or combination of elements necessary for the treatment to have benefit. For example, a study might compare Prolonged Exposure therapy with and without homework or in vivo exposure. Studies that compared exposure plus cognitive restructuring with exposure alone (Bryant, Moulds, Guthrie, Dang, & Nixon, 2003; Foa et al., 2005) also may be characterized as dismantling designs, although they have a key feature of an additive design: the combination of techniques that have already been shown to work. In an additive design, however, the effective techniques are combined and then compared with each alone, as in the study by Foa et al. (1999) that compared exposure, stress inoculation training, and their combination. In a parametric design, the amount of the active element, e.g., length of exposure sessions or depth of emotional processing, is manipulated from lesser to greater.

The fourth class of designs involves the comparison of two active, but different treatments. The comparisons may involve different types of psychotherapy (Brom, Kleber, & Defares, 1989; Resick, Nishith, Weaver, Astin, & Feuer, 2002) or drugs (Petrakis et al., 2006). Comparisons also may involve different modalities. van der Kolk and colleagues (2007) recently reported a trial comparing fluoxetine with Eye Movement Desensitization and Reprocessing (EMDR) as well as placebo. Borkovec (1993) argues that comparative studies are of extremely limited value because the many differences among various treatments prevent investigators from ruling out serious threats to validity. The problem is particularly serious in the case of studies that compare medication and psychotherapy because even nonspecific factors cannot be controlled. Studies such as van der Kolk et al.'s are nevertheless important from a practical standpoint, especially if they represent valid choices that patients or providers would make in practice.

Choosing a comparison condition. The choice of the comparison condition affects the inferences that may be drawn from a study. To understand this, it is important to remember that treatments have both nonspecific therapeutic elements, such as attention and support, as well as specific elements, such as imaginal or in vivo exposure. Change also occurs due to factors other than treatment, such as the natural history of the disorder, or to factors that are part of research, but not treatment, such as assessment. Figure 1 hypothetically illustrates which of these sources of variation occur in each comparison condition. Patients on a waiting list experience only change not due to the treatment under investigation. Patients in a nonspecific comparison group also experience change due to nonspecific therapeutic factors. Patients in an active treatment group experience change due to these two sources and change due to the specific elements of the treatment they received. As noted previously, the use of a nonspecific comparison design in CSP #494 allowed us to conclude that the benefits of Prolonged Exposure therapy extend beyond the benefits of psychotherapy, although we could not determine the mechanism by which this occurs because we did not systematically manipulate the elements of Prolonged Exposure therapy.

Figure 1 also shows that the choice of a comparison condition affects the size of the treatment effect itself. This is illustrated in Figure 2 with data from a literature review

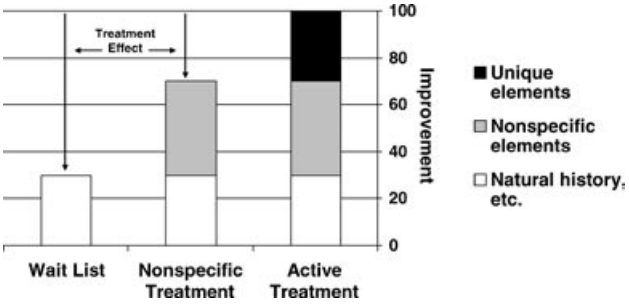

Figure 1. Hypothetical example of elements of change in wait-list, nonspecific, and active treatments. Vertical arrows indicate the effect size observed in wait-list and nonspecific groups relative to an active treatment.

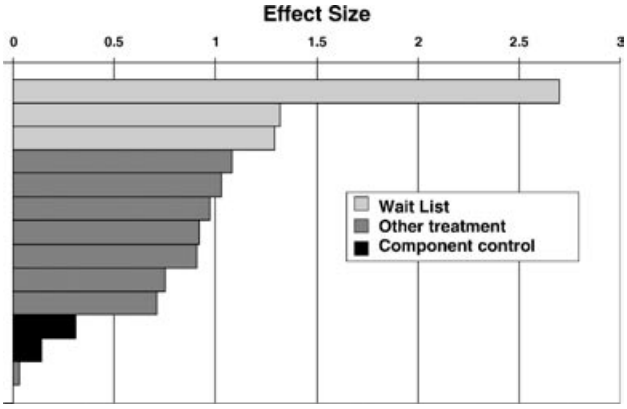

Figure 2. Effect sizes for studies of Eye Movement Desensitization and Reprocessing (EMDR) treatment (Chemtob et al., 2000, pp. 144–146, Table 7.1). Data are the largest posttreatment effect size reported for each study. For studies with more than two treatment groups, the largest effect size for each between-groups comparison is reported. Other treatment includes supportive counseling, relaxation, active listening, and imagery habituation training.

on EMDR treatment (Chemtob, Tolin, van der Kolk, & Pitman, 2000). The effect sizes ranged from 0.03 to 2.70. This variability can make it seem impossible to draw inferences about a treatment. However, the figure shows systematic differences in effect size depending on the type of comparison group. The largest effect sizes were observed in studies that employed a wait-list control, whereas two of the three smallest effect sizes were observed in dismantling studies that attempted to determine whether eye movements are necessary for patients to benefit from EMDR. By considering the type of comparison group used in each study, one would be able to conclude based on this evidence that EMDR is an effective treatment with specific benefits beyond the benefits of treatment in general, but that eye movements are not necessary for EMDR to be effective.

Table 1 presents guidelines for choosing a comparison condition based on the question one is trying to answer and the level of evidence for a given treatment. Wait-list designs are appropriate when there are no or few controlled studies because they help answer the question of whether a treatment has benefit. If there is evidence from several good wait-list controlled studies, we know that the treatment works. The question to ask at this point is whether the treatment has specific therapeutic benefits. Once this is known, dismantling, additive, or parametric designs may be appropriate if it is desirable and feasible to understand mechanism. For example, Resick et al. (2007) recently conducted a dismantling study of Cognitive Processing Therapy (CPT), comparing it to each of its two components alone, written exposure and cognitive therapy. Although CPT had not been compared to a nonspecific treatment, it had been compared and found to be equivalent to Prolonged Exposure therapy, which does have specific benefits (e.g., Bryant et al., 2003; Foa et al., 1999; Schnurr et al., 2007). It is useful to know whether both components of CPT are necessary because they could be used separately, and perhaps because the duration of treatment could be shortened if patients could derive sufficient benefits from only one component. Comparisons between active treatments should be performed when it is clear that the treatments are effective and there is a logical or substantive reason to compare them, e.g., one treatment costs more or is more difficult to deliver. van der Kolk et al.'s (2007) study is justified because there are no conclusive data to help patients and providers address a key question: Should the patient with PTSD take medication or engage in psychotherapy?

Table 2 presents recommendations for sample sizes that would yield adequate statistical power in each type of design. The table is meant to show that sample size needs rise steeply from a wait-list design to designs that employ active comparison groups, as the expected effect size progressively decreases. These are rough estimates only. Actual needs will vary based on the relative effectiveness of the treatments being compared. A case in point is the recent study by van der Kolk et al. (2007), which found differences between two active treatments based on a much smaller sample than recommended here. Needs also are affected by design strategies such as the adjustment for pretreatment values of the outcome and other covariates, which can increase statistical power (Lipsey, 1990). However, the point to remember—whether one is designing a

Table 1. Guidelines for Choosing a Comparison Condition

Comparison condition	Question	Existing evidence
Wait-list	Does the treatment have benefit?	No/few controlled studies
Nonspecific comparison or treatment as usual	Is the effect greater than the effect of simply going to therapy or getting usual care?	Good wait-list controlled studies
Combination of active elements (dismantling, additive, and parametric designs)	Why does the treatment work? What are the active ingredients?	Good nonspecific comparison studies, and it makes sense to isolate active ingredients
Other active treatment	Is treatment A better (in some way) than treatment B?	Good evidence for all treatments, and a good reason to compare treatments, e.g., cost-effectiveness

study or reading the results of one that found no differences between treatments—is how large a sample may need to be to yield adequate power for various types of designs.

Equating a Comparison Condition

Structural differences between treatment conditions on factors such as number and length of sessions, format (individual or group), type of therapist, homework, and credibility can provide alternative explanations for what appears to be a treatment effect. These factors can even influence the magnitude of the effect. Baskin, Tierney, Minami, and Wampold (2003) categorized 21 clinical trials in terms of whether the structure of the comparison group was equivalent or inequivalent to the structure of the treatment group. The effect size was significantly larger in studies with an inequivalent structure ($d = 0.47$) than in studies with an equivalent structure ($d = 0.15$).

How much should treatment conditions be equated? The answer depends on which differences could plausibly (not merely possibly) bias outcomes. Baskin et al. (2003) recommend equating treatments on number and length of sessions, format (individual vs. group), and therapist skill and training. The goal is to ensure that a comparison treatment controls for nonspecific factors, although obviously, structural factors might be intentionally manipulated in some studies, e.g., to compare individual and group formats. Baskin et al. also recommend that a comparison condition (which they refer to as a "placebo") should allow patients to discuss the problem for which they are being treated, e.g., that trauma survivors in a comparison treatment should be allowed to talk about their trauma. The example is not as straightforward as it appears. A present-centered comparison group that provides a credible rationale for focusing on current problems caused by a trauma instead of the trauma itself (e.g., McDonagh-Coyle et al., 2005; Schnurr et al., 2003, 2007) is arguably equivalent on this dimension to a trauma-focused treatment, which provides a credible rationale for focusing on the trauma. However, the equivalence could be compromised in the absence of a rationale that promoted credibility for both treatments.

Judgment is required when deciding how to equate a comparison treatment to the treatment under investigation. One helpful principle is to control to the point of "reasonableness," i.e., a reasonable person would accept

Table 2. Suggested Guidelines for Estimating Effect Size and Sample Size in a Two-Arm Trial as a Function of a Comparison Group

Comparison group	Expected effect size (Cohen's d)	N per group
Wait-list	Large ($d = .80$)	26
Nonspecific/treatment as usual	Medium ($d = .50$)	64
Component control, additive, parametric, active	Small ($d = .20$)	393

Note. Sample size estimates assume desired power = .80, $p = .05$, two-tailed (from Cohen, 1988, Table 2.4.1, p. 55).

differences between treatments as true differences. Imagine that a trauma-focused therapy consisting of 10 90-minute sessions is found to be more effective than 10 sessions of usual care, which (in this hypothetical and many actual settings) are 60 minutes long. Patients assigned to trauma-focused therapy would receive 50% more treatment than patients assigned to usual care—too much of a difference to rule out the amount of treatment as an alternative explanation. However, treatment integrity should be preserved. If patients in usual care never receive homework, which is often assigned in cognitive–behavioral treatment, assigning homework in usual care could impair the ecological validity of the treatment. This might be acceptable depending on the particular goal of the study, the level of evidence for the cognitive–behavioral treatment, and knowledge about the extent to which homework contributes to the efficacy of the cognitive–behavioral treatment. Applying the reasonableness criterion can help resolve design questions like this: Would a reasonable person actually attribute a meaningful difference between treatments to the inclusion of homework?

In CSP #494, Prolonged Exposure and Present-Centered Therapies were equated on the number (10) and duration (90-minute) of sessions, format (individual), use of a manual by therapists, psychoeducation in the initial sessions, provision of a credible rationale (that differed between treatments), and provision of homework (that also differed). The treatments did not differ in patient satisfaction or in therapist adherence and competence; patient satisfaction was even nonsignificantly higher in the present-centered group (Schnurr et al., 2007).

Assigning Therapists to Conditions

Therapists differ in terms of factors such as skill, enthusiasm for a particular technique, and warmth. Differences between therapists on these factors can affect treatment outcome, independent of the effects of the treatment. Such effects are typically indexed by the intraclass correlation (ICC), defined as the ratio of variance due to some clustering variable (such as therapists) relative to the total variance (both the cluster and error); sometimes the ratio is presented as the percentage of variance accounted for by the clustering variable by moving the decimal point. A meta-analysis of therapist effects (Crits-Cristoph et al., 1991) computed therapist effects from 15 studies that included a total of 27 treatment conditions and 141 therapists. On average, therapists accounted for 8.6% of the variance in these studies. The range across studies was substantial: 0.0% to 48.7%. Multivariate analysis showed that use of a manual and greater experience were associated with smaller therapist effects.

Therapist effects should be considered when determining how therapists are assigned to the treatment they deliver. The same therapists may administer all treatments or different therapists may administer each treatment. There are advantages and disadvantages of each approach (Schnurr, Friedman, Lavori, & Hsieh, 2001; Schnurr et al., 2005). Although it may seem that having a therapist deliver all treatments would control for therapist effects, this strategy does not guarantee control. A therapist may be more skilled in delivering one treatment, or may prefer one treatment to another. A therapist also may find it difficult to keep the treatments distinct if delivering more than one. However, Schnurr et al. (2005) suggest that these risks may be necessary in trials in which there are few therapists and recommend that therapists are thus carefully supervised to detect and correct problems if they arise.

Randomizing therapists to deliver one treatment can be used as an alternative strategy to control therapist effects in trials with more than a few therapists. For example, the 52 therapists in CSP #494 were randomly assigned to deliver Prolonged Exposure or Present-Centered Therapy. Randomizing therapists affords the same benefits as randomizing patients: individual differences are randomly distributed across conditions. However, randomization requires therapists willing and able to deliver any of the treatments being studied, and, like the other strategies, it offers no guarantees. Therapists who believed they would accept either treatment may find this not to be the case. Although some therapists in CSP #494 who were assigned to Present-Centered Therapy initially expressed concerns about delivering a lesser therapy, these concerns were resolved in all but one case. In addition, one potential complication of having

different therapists deliver each treatment is that a patient cannot be randomized unless there is an available therapist in each condition. This presented scheduling problems in CSP #494 given the busy schedules of our therapists; one eligible patient actually could not be treated because these problems could not be resolved.

Having different therapists deliver each treatment is also appropriate if experts are needed to deliver the different treatments, as in the National Institute of Mental Health Treatment of Depression Collaborative Research Program (Elkin et al., 1989), which contrasted cognitive–behavioral therapy with interpersonal therapy (and also antidepressant medication). No matter what assignment strategy is used, careful training and supervision are necessary to try to minimize therapist effects. Moreover, therapist effects should be accounted for in sample size projection (see the section below on group therapy) and the therapist should be treated as a fixed or random factor (depending on the study design) in data analysis (cf. Elkin, Falconnier, Martinovich, & Mahoney, 2006; Kim, Wampold, & Bolt, 2006; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007). In CSP #494, therapist effects were small (intraclass $r = .05$), but failing to adjust for even such a low correlation would have inflated the treatment effect.

Manualization

Manuals are an essential component of quality control in psychotherapy research (Borkovec, 1993; Foa & Meadows, 1997). Manuals facilitate consistent treatment delivery, replication, and dissemination. They also enhance training, supervision, and fidelity monitoring. Borkovec suggests that manuals should contain the theoretical background for a treatment, the rationale for patients, and the procedures to be used. Detail is essential. Manuals should have session-by-session guidelines, along with specific prompts and suggestions to be used to implement the procedures and address patient issues that hinder implementation of the protocol. Manuals for different treatments should be as comparable as possible to equate their credibility, although it may be necessary for them to differ in length. In CSP #494, the manual for Present-Centered Therapy was shorter than the manual for Prolonged Exposure therapy because of the more detailed structure of the Prolonged Exposure therapy sessions. This is an example of how treatment integrity should be preserved when trying to equate conditions. Padding the Present-Centered Therapy manual with unnecessary content just to extend its length likely would have impaired its credibility rather than enhance it.

Training, Supervision, and Monitoring

Several additional elements are essential to quality control in psychotherapy research: training therapists how to deliver the treatment, supervising therapists during the study, and independently monitoring fidelity and competence. Checking alliance and process are types of monitoring too, although the discussion of these topics is beyond the scope of this article.

Prior to beginning a study, investigators should determine the type and amount of training to be provided before therapists are allowed to treat patients. Training may involve a combination of face-to-face instruction, telephone calls, and assigned reading. It is desirable to equate the amount and format of training across conditions, although this may not be feasible. In CSP #494, the in-person training was 5 days for Prolonged Exposure therapy therapists and 2 days for Present-Centered Therapy therapists because of the differential amount of new material required in each condition. Training also typically involves treating practice cases under expert supervision, usually to some fixed number of cases or standard of mastery. Therapists in CSP #494 treated one to two practice cases.

Ongoing supervision is also used to ensure that therapists maintain proficiency and manual adherence throughout a study. It may be based on videotape, audiotape, or therapist report and vary in format (individual vs. group; phone or in-person) and frequency. Supervision should be more rigorous in efficacy studies, and less so in effectiveness studies, which should use strategies that resemble how supervision would be conducted in practice. In CSP #494, videotapes of all sessions were viewed by experts who provided telephone supervision weekly for each therapist's first 4 patients (after their practices cases) and every other week

for the therapist's additional patients. Use of a supervision checklist is recommended, particularly if there are many supervisors.

Monitoring by an expert independent of the training and supervision process is an extremely valuable source of information about therapists' adherence to the treatment protocol, skill, and other behaviors that could affect treatment outcome. Monitoring provides a check on the extent to which training and supervision succeeded and on the comparability of adherence and therapist factors across treatments. It should be based on session videotapes, although audiotapes are acceptable. Typically, no matter how a random sample of sessions is rated, the sampling percentage should depend on the number of sessions judged by the investigators to be necessary for the monitor to see a representative sample of treatment. In CSP #494, the fidelity monitor rated 11.7% of the videotapes ($n = 269$).

The monitor should use a scale that assesses therapist performance on (a) elements unique to a particular treatment, (b) necessary but not unique to a particular treatment, and (c) proscribed in a particular treatment. The fact that treatments vary in their unique elements and in the proscribed elements means it is essential for a scale to be constructed so that it is possible to compare different treatments, e.g., with global ratings of competence and adherence, or with scales that can be converted to aggregate percentages. Among the items rated by the monitor in CSP #494, there was a 5-point scale (1 = *poor*, 5 = *excellent*) for therapists' competence and adherence to essential manual elements that were (a) unique to that approach, and (b) not unique to that approach. Proscribed elements were rated present/absent and were converted to a percentage for each tape.

Additional Treatment

Many patients screened for entering trials are already undergoing treatment. In CSP #494, over 60% of patients were receiving psychotherapy and 75% were taking medication. The vast majority of psychotherapy studies in the trauma field allow some form of concurrent treatment (e.g., Cloitre et al., 2002; Foa et al., 1999, 2005; McDonagh-Coyle et al., 2005; Monson et al., 2006; Resick et al., 2002; Schnurr et al., 2003, 2007). Typically, patients must be on a stable medication regimen (e.g., 2–3 months prior) and discontinue psychotherapy for PTSD and sometimes for related disorders. Medication change is sometimes formally allowed (e.g., Resick et al., 2002; Schnurr et al., 2003, 2007), although all patients should be included in ITT analysis regardless of whether their medication changes.

Considerations of both ethics and feasibility should determine whether patients are required to discontinue concurrent treatments. Stricter control is needed in efficacy studies, so it is desirable to allow only treatments that are necessary for ethical reasons. Less control is needed in effectiveness studies, in which it is acceptable to allow additional treatments that do not interfere with study treatment, e.g., support groups. Nevertheless, it is prudent to require medications to be stabilized before study, and to discourage change unless clinically necessary. Concurrent PTSD treatment probably should be disallowed or at least strongly discouraged on ethical grounds, as it is not good practice to have two noncollaborating therapists treating a patient for the same condition. Concurrent therapy should be assessed to determine whether it differs between treatments and/or modifies the treatment effect.

In CSP #494, patients were allowed to be on medication, although a stable regimen of 2 months before entry was required. Unnecessary change was discouraged. Patients were allowed to attend self-help groups, have brief check-ins with an existing therapist, and receive psychotherapy for problems that were not PTSD-related. Treatment conditions were comparable in terms of all measures of concurrent therapy before and during treatment, except the Present-Centered Therapy group was more likely than the Prolonged Exposure therapy group to receive a new medication or an increased dose of medication during study treatment. This could have attenuated the treatment effect (if the changes improved patients' symptoms); however, analyses showed that medication did not modify the treatment effect. One strategy to ensure that medication change is clinically indicated is to have medications prescribed and monitored by a member of the study team (e.g., Resick et al., 2002). It is critically important for

this person to be blind to the type of psychotherapy the patient is receiving.

Group-Based Treatments

Psychotherapy interventions may be delivered in a group format as well as individual format. As discussed previously, the outcomes of patients treated by a given therapist may be correlated. Outcomes of patients treated in a group format may be similarly correlated (Baldwin, Murray, & Shadish, 2005; Burlingame, Kircher, & Honts, 1994; Schnurr et al., 2001). Such effects, like therapist effects, can be indexed as an ICC. For example, Schnurr et al. (2003) reported an ICC of .04 for total PTSD severity in their group psychotherapy study. Creamer, Morris, Biddle, and Elliot (2000) reported even larger ICCs for group treatment: .13 for PTSD severity, .12 for intrusion, .10 for avoidance, and .13 for arousal.

Failing to account for such correlations can inflate Type I error by underestimating the standard error (cf. Burlingame et al., 1994). The consequences can be serious. Baldwin and colleagues (2005) reanalyzed findings from 33 studies of group therapy on the list of empirically supported treatments produced by the American Psychological Association's Task Force on Promotion and Dissemination of Psychological Procedures (1998). The original analyses were corrected for various assumptions of within-group correlation and for degrees of freedom, which should be based on the number of groups and not the total number of subjects. After correction for degrees of freedom, only 50.8%–68.2% of the previously significant findings were still significant, depending on assumptions made about the tests. Also assuming an ICC = .05, only 35.2%–43.3% of the tests remained significant.

In somewhat more technical terms, the problem caused by group clustering is called the *variance inflation factor* (VIF): $\text{VIF} = 1 + (n \text{ per group} - 1)\text{ICC}$. This is the additional variance added to the treatment effect that needs to be accounted for in the analyses by treating group as a random factor with participants nested within groups. Such adjustments are rare in the field of trauma research (see, e.g., Creamer et al., 1999; Schnurr et al., 2003, for exceptions).

Group clustering also should be accounted for in sample size projection (Baldwin et al., 2005; Schnurr et al., 2001). In a two-group design, 128 participants ($n = 64$ per condition) would be needed to have .80 power to detect a medium effect of $d = 0.50$ at $\alpha = .05$ (two-tailed; Cohen, 1988). Even a small ICC inflates sample size needs substantially. For example, if ICC = .05, the needed N increases to 170 for groups of 5 members and 220 for groups of 10 members (Baldwin et al., 2005). Because the unit of analysis is the group, fewer participants are needed if distributed across more (smaller) groups.

CONCLUSIONS

Although psychotherapy trials have unique features that can create challenging methodological questions, it is possible to conduct valid research on psychotherapy. However, researchers need increased knowledge to help them safely navigate between what can seem like "rocks and hard places" in study design. Reviewers need increased knowledge to offer optimal feedback to both researchers and as do journal editors and funding agencies. Readers also need increased knowledge to help them evaluate the results of psychotherapy studies and use the findings appropriately, whether to decide about engaging in a particular treatment or making policy decisions for a hospital system.

This article is an attempt to meet these diverse needs by providing information about significant methodological issues in psychotherapy outcome research. Additional support for this goal comes from a recent initiative to expand the CONSORT standards for reporting clinical trials of nonpharmacological interventions such as psychotherapy and surgery (Boutron, Moher, Altman, Schultz, & Ravaud, 2007, for the CONSORT Group). CONSORT—Consolidated Standards on Reporting Trials—was developed to facilitate complete and accurate reporting of clinical trials (Moher, Schultz, & Altman for the CONSORT Group, 2001). The CONSORT standards require a figure depicting participant flow from screening through analysis, and a 22-item checklist prepared by authors for

reviewers and editors to use in finding information about all 22 specific elements, which include items covering design, analysis, and writing. Although it was decided that a new CONSORT statement would not be required, the language of 11 items was modified to capture the unique information that needs to be conveyed, such as adherence to a therapy manual. Improving the reporting of trials in this way can enhance not just the quality of information to be derived from the published literature, but also the quality of studies that are designed with the standards in mind.

In his classic book *Great Expectations* (1861/2006), Dickens recommends, "Take nothing on its looks; take everything on evidence. There's no better rule" (p. 327). His comment is prescient of recent efforts to promote empirically supported treatment, such as the American Psychological Association's Task Force on Promotion and Dissemination of Psychological Procedures (1998). These efforts require high-quality evidence. Yet we must consistently strive to study ways to promote dissemination as well, to ensure that treatments are implemented in practice. Our research needs to keep both goals in mind if we are to follow Dickens' advice.

REFERENCES

American Psychological Association's Task Force on Promotion and Dissemination of Psychological Procedures. (1998). Update on empirically supported therapies II. Clinical Psychologist, 51, 3–16.

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. Journal of Consulting and Clinical Psychology, 73, 924–935.

Baskin, T. W., Tierney, S. C., Minami, T., & Wampold, B. E. (2003). Establishing specificity in psychotherapy: A meta-analysis of structural equivalence of placebo controls. Journal of Consulting and Clinical Psychology, 71, 973–979.

Borkovec, T. D. (1993). Between-group therapy outcome research: Design and methodology. NIDA Research Monograph, 137, 249–289.

Boutron, I., Moher, D., Altman, D. G., Schultz, K. F., & Ravaud, P. for the CONSORT Group (2007). Reporting of nonpharmacological treatment interventions: An extension of the CONSORT Statement. Manuscript submitted for publication.

Brom, D., Kleber, R. J., & Defares, P. B. (1989). Brief psychotherapy for posttraumatic stress disorders. Journal of Consulting and Clinical Psychology, 57, 607–612.

Bryant, R. A., Moulds, M. L., Guthrie, R. M., Dang, S. T., & Nixon, R. D. V. (2003). Imaginal exposure alone and imaginal exposure with cognitive restructuring in treatment of posttraumatic stress disorder. Journal of Consulting and Clinical Psychology, 71, 706–712.

Burlingame, G. M., Kircher, J. C., & Honts, C. R. (1994). Analysis of variance versus bootstrap procedures for analyzing dependent observations in small group research. Small Group Research, 25, 486–501.

Chemtob, C., Tolin, D., van der Kolk, B., & Pitman, R. (2000). Eye Movement Desensitization and Reprocessing. In E. B. Foa, T. M. Keane, & M. J. Friedman (Eds.), Effective treatments for PTSD (pp. 139–154). New York: Guilford Press.

Cloitre, M., Koenan, K. C., Cohen, L., & Han, H. (2002). Skills training in affective and interpersonal regulation followed by exposure: A phase-based treatment for PTSD related to childhood abuse. Journal of Consulting and Clinical Psychology, 70, 1067–1074.

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

Cook, T. D., & Campbell, T. D. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Creamer, M., Morris, P., Biddle, D., & Elliott, P. (1999). Treatment outcome in Australian veterans with posttraumatic stress disorder: A cause for cautious optimism? Journal of Traumatic Stress, 12, 545–558.

Crits-Cristoph, P., Baranckie, K., Kurcias, J. S., Beck, A., Carroll, K., Perry, K., et al. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. Psychotherapy Research, 1, 81–91.

Dickens, C. (2006). Great expectations. Ann Arbor, MI: Ann Arbor Media. (Original work published 1861)

Elkin, I., Falconnier, L., Martinovich, Z., & Mahoney, C. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. Psychotherapy Research, 16, 144–160.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., et al. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program:

General effectiveness of treatments. Archives of General Psychiatry, 46, 971–982.

Foa, E. B., Dançu, C. V., Hembree, E. A., et al. (1999). A comparison of Prolonged Exposure, stress inoculation, and their combination for reducing posttraumatic stress disorder in female assault victims. Journal of Consulting and Clinical Psychology, 67, 194–200.

Foa, E. B., Hembree, E. A., Cahill, S. P., Rauch, S. A. M., Riggs, D. S., Feeny, N. C., et al. (2005). Randomized trial of Prolonged Exposure for posttraumatic stress disorder with and without cognitive restructuring: Outcome at academic and community clinics. Journal of Consulting and Clinical Psychology, 73, 953–964.

Foa, E. B., & Meadows, E. (1997). Psychosocial treatments for posttraumatic stress disorder: A critical review. Annual Review of Psychology, 48, 449–480.

Foa, E. B., & Rothbaum, B. O. (1998). Treating the trauma of rape: Cognitive-behavioral therapy for PTSD, New York, Gullford.

Goetghebeur, E., & Loeys, T. (2002). Beyond intention to treat. Epidemiologic Reviews, 24, 85–90.

Kim, D.-M., Wampold, B. E., & Bolt, D. M. (2006). Therapist in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. Psychotherapy Research, 16, 161–172.

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. Controlled Clinical Trials, 21, 167–189.

Lipsey, M. W. (1990). Design sensitivity: Statistical power for experimental research. Newbury Park, CA: Sage.

Lutz, W., Leon, C. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth-curve approach. Journal of Counseling Psychology, 54, 32–39.

McDonagh-Coyle, A., Friedman, M. J., McHugo, G. J., Ford, J., Sengupta, A., Mueser, K., et al. (2005). Randomized trial of cognitive behavioral therapy for chronic PTSD in adult female childhood sexual abuse survivors. Journal of Consulting and Clinical Psychology, 73, 515–524.

McFall, M., Saxon, A. J., Thompson, C. E., Yoshimoto, D., Malte, C., Straits-Troster, K., et al. (2005). Improving the rates of quitting smoking for veterans with posttraumatic stress disorder. American Journal of Psychiatry, 162, 1311–1319.

Moher, D., Schultz, K., & Altman, D. for the CONSORT group. (2001). The CONSORT Statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. Journal of the American Medical Association, 285, 1987–1991.

Monson, C. M., Schnurr, P. P., Resick, P. A., Friedman, M. J., Young-Xu, Y., & Stevens, S. P. (2006). A randomized controlled trial of Cognitive Processing Therapy for veterans with military-related posttraumatic stress disorder. Journal of Consulting and Clinical Psychology, 74, 898–907.

Moseley, J. B., O'Malley, K., Petersen, N. J., Menke, T. J., Brody, B. A., Kuykendall, D. H., et al. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee. New England Journal of Medicine, 347, 81–88.

Peto, R., Pike, M., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., et al. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. British Journal of Cancer, 35, 1–39.

Petrakis, I. L., Poling, J., Levinson, C. M., Nich, C., Carroll, K. M., Ralevski, E., et al. (2006). Naltrexone and disulfiram in patients with alcohol dependence and comorbid posttraumatic stress disorder. Biological Psychiatry, 60, 777–783.

Resick, P. A., Galovski, T., Uhlmansiek, M. H., Scher, C. S., Clum, G., & Young-Xu, Y. (2007). A randomized clinical trial to dismantle components of Cognitive Processing Therapy for posttraumatic stress disorder in female victims of interpersonal violence. Manuscript submitted for publication.

Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of Cognitive Processing Therapy with Prolonged Exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. Journal of Consulting and Clinical Psychology, 70, 867–869.

Rounsaville, B. J., Carroll, K. M., & Onken. L. (2001). A stage model of behavioral therapies research: Getting started and moving on from Stage 1. Clinical Psychology: Science and Practice, 8, 133–142.

Schnurr, P. P., Friedman, M. J., Engel, C. C., Foa, E. B., Shea, M. T., Chow, B. K., et al. (2007). Cognitive–behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. Journal of the American Medical Association, 297, 820–830.

Schnurr, P. P., Friedman, M. J., Engel, C. C., Foa, E. B., Shea, M. T., Resick, P. M., et al. (2005). Issues in the design of multisite clinical trials of psychotherapy: CSP #494 as an example. Contemporary Clinical Trials, 26, 626–636.

Schnurr, P. P., Friedman, M. J., Foy, D. W., Shea, M. T., Hsieh, F. Y., Lavori, P. W., et al. (2003). A randomized trial of trauma focus group therapy for posttraumatic stress disorder: Results from a Department of Veterans Affairs Cooperative Study. Archives of General Psychiatry, 60, 481–489.

Schnurr, P. P., Friedman, M. J., Lavori, P. J., & Hsieh, F. Y. (2001). Design of VA Cooperative Study #420: Group treatment of PTSD. Controlled Clinical Trials, 22, 74–88.

Seligman, M. E. (1995). The effectiveness of psychotherapy. The Consumer Reports study. American Psychologist, 50, 965–974.

Tunis, S. R., Stryer, D. B., & Clancy, C. M. (2003). Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. Journal of the American Medical Association, 290, 1624–1632.

van der Kolk, B. A., Spinazzola, J., Blaustein, M. E., Hopper, J. W., Hopper, E. K., Korn, D. L., et al. (2007). A randomized clinical trial of Eye Movement Desensitization and Reprocessing (EMDR), fluoxetine, and pill placebo in the treatment of posttraumatic stress disorder: Treatment effects and long-term maintenance. Journal of Clinical Psychiatry, 68, 37–46.