32

TREATMENT OUTCOME RESEARCH

MEGAN E. SPOKAS

THOMAS L. RODEBAUGH

RICHARD G. HEIMBERG

Clinical psychology can be thought of as a science with two primary aims: understanding the nature of psychological distress and developing methods to relieve that distress. Treatment outcome studies play a vital role in both those endeavors, and they are a commonly used collection of methodologies in clinical psychology research. The main thrust of this approach involves evaluating the efficacy of interventions and investigating the mechanisms by which effective interventions produce change. The findings allow clinicians a greater understanding of how psychological distress can be reduced, and the implications of the findings inform their understanding of the nature of psychological distress. Knowledge of the various study designs used to assess treatment outcomes should be helpful not only to individuals who wish to design treatment studies but also to those who evaluate those studies. Clinicians who wish to evaluate whether a treatment method may be useful for their clients will be able to evaluate available studies only to the extent that they understand the logic of specific study designs. To shed light on this area of study, we first review the various types of treatment outcome study designs and then discuss some of the many considerations for this type of investigation.

TYPES OF TREATMENT OUTCOME DESIGNS

No-Treatment Control Group Designs

The first stage of treatment development often involves an open pilot study, or simply evaluating change in participants' symptoms across time. However, this design does not allow the researcher to attribute positive outcomes to the treatment, as no potentially competing explanations for a positive result have been ruled out. Thus, an important next step is comparing the treatment package to the passage of time. In the simplest form of this type of design, participants who are randomly assigned to receive treatment are compared to those randomly assigned to a no-treatment condition, and participants in both conditions are assessed before and after an equivalent period of time elapses. This is the most liberal group comparison in treatment outcome research, but when developing new treatments, it is a useful first step to determine whether the new treatment is more efficacious than no treatment at all. Results that indicate a difference suggest that further investigation or refinement of the fledgling treatment is warranted. The use of a no-treatment control group also provides important information about how presenting problems change over

time without treatment. For example, Posternak and Miller (2001), in a meta-analysis of participants who were on a wait-list for treatment of depression, found that 19.7% of participants showed symptom reduction consistent with a successful trial of treatment with antidepressant medication. Psychological distress or disorder may remit or change naturally over time, and it is important to control for any naturally occurring factors that can affect outcome. Other threats to internal validity, such as maturation and effects of assessment, may also be controlled with this type of design.

The ethical complications involved in a study comparing an active treatment to a strict no-treatment control group are obvious; withholding treatment from someone presenting for treatment may be questionable under many circumstances. If a no-treatment control group appears ethically justifiable, we would recommend particular attention be paid to the assessment and withdrawal of participants from the study if the target problem becomes more distressing or if other problems, such as depression or demoralization, develop. Treatment can then be provided to the person without the constraints imposed by study participation. Both ethical and research considerations provide motivation to limit the duration of no-treatment conditions. With longer time frames, participants may be more likely to drop out of the study. In addition, the longer the no-treatment interval, the greater the potential ethical culpability of the researcher. These concerns may be either compounded or alleviated depending upon the course and severity of the presenting problem being treated. If it is likely that the no-treatment control group will significantly increase in severity of symptoms, the researcher bears increased responsibility for demonstrating that a strict no-treatment control condition is, overall, the best comparison condition to use. In fact, we find it difficult to think of any research situation in which a pure no-treatment control condition seems to be the best possible control condition. Instead, other forms of no-treatment control conditions, such as wait-list control groups, may provide many of the advantages and fewer of the disadvantages of no-treatment control groups.

Wait-List Control Group Designs

Some of the ethical concerns associated with a no-treatment control design are addressed in a wait-list control group design. In this design, all participants receive an offer of treatment, but some receive it immediately after enrolling in the study, while others are required to wait a specified amount of time, typically the same time that is devoted to the administration of treatment to those who receive it immediately. This design ensures that all participants ultimately receive some type of active treatment, while also allowing the investigator to compare the progress of participants in the active treatment to those receiving no treatment. However, in this design, the long-term effects of maturation or the natural course of the problem cannot usually be assessed and controlled because treatment is administered as soon as the waiting period is over. Another advantage of this design is that the wait-list control group not only allows for a comparison between the effects of receiving treatment versus no treatment, but ultimately there are more participants receiving the treatment of interest. If the wait-list group is assessed again after they have received treatment, a replication of the findings for the immediate treatment group can be conducted, and a larger sample size is available for secondary analysis, such as the determination of predictors of successful outcomes.

Attention Control Group Designs

When using a no-treatment or wait-list control group design, it can be difficult to determine if differences between groups in post-treatment (vs. post-wait) symptoms are due to the treatment or to other factors associated with the research study. For example, having repeated contact with health care providers, taking action toward managing the presenting problem, gaining more knowledge about the problem in the context of the study, and gaining increased hope or expectations of improvement may contribute to observed changes. These factors are commonly referred to as "nonspecific" because they are present in most therapies. As Borkovec (1994) notes, "Researchers are interested ultimately in whether or not a specific therapy contains active ingredients for change in addition to these common ingredients" (p. 256).

An attention control group can be used to test the rival hypothesis that attention during treatment and/or participants' expectations about treatment are responsible for observed changes. In this type of study, investigators

design a control condition that provides participants with an experience similar to that of the treatment group, equating the nonspecific factors across conditions without actually providing the theoretically active treatment ingredient. For example, the control condition may involve meeting with a therapist for the same number of sessions as in the active treatment, as well as more specific elements, such as education about the presenting problem, time spent talking about the problem, or other procedures that are equivalent in duration to the treatment group. For example, in a study assessing the effects of cognitive preparation when using video feedback for reducing social anxiety symptoms, Rodebaugh (2004) compared the active cognitive preparation condition to an attention control condition. Ninety-five speech-anxious college students were randomized to one of two conditions. Both conditions involved the same experiences (giving three speeches while being videotaped), the same assessments and amount of contact with the experimenter (through yoking), and similar presentations of a treatment rationale: "Most people look better giving speeches than they believe they do" (p. 1440). However, participants in the treatment condition then engaged in cognitive preparation: first imagining and visualizing what they would look like in the video and then watching the video of themselves as if they were watching a stranger. In contrast, the control group engaged in (theoretically) neutral imagery tasks. Group differences in self-perception were found after treatment. The use of an attention control condition such as this one potentially allows the experimenter to rule out the rival hypothesis that the difference was due to differences in attention from the experimenter and/or expectations on the part of the participants.

Most full-scale treatment studies do not use attention control groups because of the difficulties involved in developing a control condition that is equal to the active treatment along the dimensions of credibility of treatment rationales and the ability of the treatment of control condition to generate expectations for positive treatment outcome. One study that did so was conducted by Heimberg et al. (1990), who compared cognitive-behavioral group therapy (CBGT) for social anxiety disorder to an educational-supportive group psychotherapy (ES) control designed for the study. Pilot testing of the therapy rationales revealed the educational-supportive intervention to be equally credible to cognitive-behavioral therapy, and ratings of credibility and positive outcome expectations provided by study participants showed no differences between the CBGT and ES conditions. Nevertheless, CBGT was substantially more efficacious than ES. Heimberg et al. (1998) replicated this finding. They also demonstrated that the effects of ES were approximately the same as those demonstrated by a pill placebo condition.

However, the best-laid plans of therapy researchers can go awry. Butler, Cullington, Munby, Amies, and Gelder (1984) examined the efficacy of cognitive-behavioral treatment for social anxiety disorder compared to that of *in vivo* exposure alone. To make sure that the exposure alone condition matched the cognitive-behavioral treatment in numbers of sessions and other criteria, Butler et al. substituted an attention-placebo procedure called "associative therapy" for the cognitive techniques in the cognitive-behavioral treatment. At the end of treatment, the cognitive-behavioral treatment was deemed more efficacious than exposure alone. However, this interpretation was clouded by the finding that it was also judged by participants at the fourth session to be more likely to produce a positive outcome.

Treatment-as-Usual Control Group Designs

Because developing an adequate attention control can be difficult, some researchers compare the treatment of interest with a treatment that individuals typically receive in the community, often referred to as "treatment as usual" (TAU). This design more adequately addresses ethical concerns about withholding treatment or not providing active treatment, because the comparison group has the freedom to pursue the treatment that they would likely receive if they sought treatment outside of the research study. Furthermore, the treatment being evaluated in this design represents an attempt to improve the outcome observed beyond that provided in real life. This design also controls for the common or nonspecific components of therapy discussed previously.

A study investigating the effects of dialectical-behavior therapy (DBT) in the treatment of suicidal adolescents (Katz, Cox, Gunasekara, & Miller, 2004) provides an example of the use of a TAU control group. DBT was administered to a group of inpatient adolescents, who were then

compared to participants on another hospital unit who received TAU (in this case, psychodynamic psychotherapy). The treatment programs were comparable in terms of amount of contact with care providers. DBT and TAU were associated with comparable reductions in depressive symptoms, parasuicidal behavior, and suicidal ideation, up to one year after discharge, whereas DBT produced a greater reduction in behavioral incidents (typically involving attempts to harm self or others) during hospitalization.

Although TAU groups are sometimes thought of as more ethical comparison conditions, especially when studying treatments for severe behavioral and emotional problems, there may also be significant problems. Spiritu, Stanton, Donaldson, and Boergers (2002) discuss the difficulties that arise when the treatment involved in the TAU condition is not directly measured and controlled. They conducted a study of treatment for adolescent suicide attempters in which participants were randomized to a problem-solving intervention designed to increase adherence with outpatient treatment or standard disposition planning (TAU). Evaluation of the TAU condition revealed that fewer than one-half of the adolescents in this condition attended six or more therapy sessions. Furthermore, the content of the sessions varied greatly, with supportive psychotherapy techniques reported by three-fourths of the sample, psychodynamic and cognitive techniques by one-half of the sample, and behavioral techniques by one-third of the sample. Because of this variability in number and type of treatment sessions received, it is difficult to determine the meaningfulness of differences between the problem-solving intervention and this TAU control condition. Proper use of this design involves detailed description and careful monitoring of the TAU, so that all participants in the control group are receiving a similar intervention.

Such concerns are clearly important in regard to the types of inferences that can be made regarding the causal mechanisms involved in a treatment, but they may be less crucial for other types of questions. If the unit of analysis is the last observation carried forward (see "Attrition and Intent-to-Treat Analyses" later in the chapter), then the comparison between TAU and another treatment is on par with a comparison between any other treatment conditions: the inability of TAU to retain clients can be taken as an indication of the shortcoming of TAU rather than as a problem with the design per se. From a public health standpoint, it can be useful to demonstrate that a treatment is better than TAU, but from a standpoint of determining the specific mechanisms involved in therapeutic change, such a comparison leaves much to be desired unless the TAU is readily definable and can effectively control for nonspecific aspects of treatment.

Dismantling Treatment Designs

When there is evidence for the efficacy of a treatment package, a researcher may conduct a study with a dismantling treatment design to determine which components of the treatment package are responsible for the observed changes. This type of design involves two or more groups who receive varying components of the original treatment package and requires that the treatment package be multifaceted such that specific components of the treatment can be identified. For example, one group may receive the entire treatment package, whereas another group may receive the package without one of the original components included. Various individual components can be isolated and studied by adding groups that either do or do not receive specific components of interest.

For example, Schmidt and colleagues (2002) were interested in studying the utility of breathing retraining (BR) as part of the treatment of panic disorder. They started with a cognitive-behavioral treatment (CBT) package that included the following components: (1) psychoeducation, (2) cognitive restructuring, (3) exposure to bodily sensations, (4) exposure to external stimuli, and (5) breathing retraining. The design involved three groups: one wait-list control group and two groups receiving CBT, one with BR and one without BR. If the CBT group with BR had demonstrated significantly more improvement, this would suggest that BR was crucial to the efficacy of the treatment package. However, the two CBT groups were equivalent in terms of improvement, which led the investigators to question the need to include breathing retraining in the treatment package.

As with all demonstrations of null results, this outcome is not as easy to interpret as results indicating an effect due to the element in question. Clinical samples are typically not large, and dismantling study designs typically involve more conditions than some of the other designs described earlier. As a result, there may be fewer

participants in each condition and less power to detect group differences. If the researcher thoroughly evaluates the constructs in question (e.g., by using a variety of valid indicators of the symptoms that are expected to change in treatment), then additional power to detect differences can be obtained through combining measures (e.g., through multivariate techniques; see "Statistical Concerns" following), increasing reliability of assessment. To clearly demonstrate a lack of meaningful difference, the researcher must carefully attend to issues of power and the comprehensiveness of assessment.

Constructive Treatment Designs

A constructive treatment design can assist in building new treatment packages. A core component can be assessed, and then additional components can be added and evaluated to see if they increase treatment efficacy. After several additional components are evaluated and then added to or discarded from the treatment, a researcher can determine the most effective and efficient package among the components evaluated. A constructive treatment design can also be used when a researcher is interested in finding ways to enhance treatment outcome by adding components to already developed treatment packages. Even when a treatment package is found to be superior to other forms of treatment, we have yet to see a response rate of 100%, leaving room for improving treatment response. For example, in a study of treatment for marijuana dependence, a voucher-based incentive program was added to a treatment involving motivational enhancement and the learning of behavioral coping skills (Budney, Higgins, Radonovich, & Novy, 2000). This combined treatment was compared to a treatment involving motivational enhancement and coping skills and one including only motivational enhancement. Compared to the other two groups, the participants in the incentive program evidenced greater durations of abstinence from marijuana over the course of the 14-week study, and a greater percentage of participants in the incentive program group were abstinent at the end of the treatment. The addition of the incentive program appeared to enhance improvement rates.

A constructive design may also involve adding an entire treatment package. For example, Barrett, Dadds, and Rapee (1996) investigated the addition of family therapy to a cognitive-behavioral treatment for anxious children. Anxious children were randomized to one of three conditions: CBT, CBT plus family therapy (CBT + FAM), or a wait-list control. Both active treatments were superior to the wait-list condition; however, at 12-month follow-up, fewer children in the CBT + FAM condition continued to meet criteria for an anxiety disorder: 5% versus 30% in the CBT group. These results suggest that the addition of another therapy program may enhance long-term outcome.

Although it is often assumed that more treatment is better, it is important to note that two treatments are not always better than one. In some cases, adding another treatment component or package may become confusing or overwhelming to participants. In other cases, adding additional elements may result in longer periods of time needed to reach the desired results because more sets of skills must be learned. Alternatively, the two components may not be easily integrated, leading to a decrease in overall effect. For example, in a study of various treatments for panic disorder (Barlow, Gorman, Shear, & Woods, 2000), the combination of CBT and medication (imipramine) was found to be less effective at follow-up than CBT without medication. This provides an example of a single treatment being more effective than a combined treatment, and it may also suggest that the addition of a treatment can interfere with the maintenance of gains that may be associated with the other treatment (Foa, Franklin, & Moser, 2002).

Both dismantling and constructive designs present problems with which the researcher must cope. When researchers wish to control for the passage of time or, more specifically, time spent interacting with a therapist, adding or subtracting treatment elements becomes an issue. The researcher must consider whether it is reasonable, for example, for a therapy with two components to be carried out in the same amount of time as a therapy with one component. If it is not, the researcher may be testing the difference between two ineffectively administered components and one effectively administered one, which is unlikely to be the theoretical question of interest. In contrast, if the researcher decides to allow one therapy to take more time than another, time spent with the therapist may be a confounding variable. This issue may be resolved by using a short and a long version of each individual treatment component. For example, a study of dating

skills training investigated two treatments: response acquisition and cognitive self-statement modification (Glass, Gottman, & Shmurak, 1976). These investigators included a control group, a group for each of the two treatments, and a group of combined treatment. In addition, to control for the longer time period needed to effectively administer the combined treatment, enhanced versions of each of the individual treatments (with treatment lengths equaling the length of the combined treatment condition) were added. Participants trained in cognitive self-statement modification evidenced significantly more improvements, which were maintained or enhanced at six-month follow-up. This study was able to control for differences in length of treatment across conditions; however, a design such as this is costly, both in resources and number of participants required.

Comparative Treatment Designs

In contrast to comparing a treatment to no treatment or a treatment believed to have only minimal or nonspecific effects, one can compare the treatment of interest to another type of treatment that differs in theoretical approach. This type of design may answer the question, "Which type of therapy is better for this particular problem?" Examples include comparisons of cognitive-behavioral therapy to interpersonal therapy (IPT) or brief psychodynamic therapy. The comparison treatments may differ greatly in terms of their proposed theoretical mechanism, or, alternatively, one approach may represent a (usually multifaceted) modification of the other.

It can be argued that comparative treatment designs are important from a public health perspective: Both providers and individuals seeking treatment are interested in knowing what treatment will be most effective in treating a particular problem. However, the scientific questions that can be answered with such designs are limited, as some authors (Borkovec & Miranda, 1999; Kazdin, 2001, 2004) have indicated. Borkovec and Miranda outline several factors that limit the results of a comparative treatment design. For example, when comparing two radically different treatment approaches, results may indicate that one treatment produces superior effects, but because it is unlikely that patients' experiences are equivalent across treatment conditions, the mechanisms by which the superior outcome was achieved is unclear. Another important limitation is the potential for differences in the quality of treatment provided. If therapists in a research study are well versed in one treatment modality but have little experience in providing the comparison treatment, differences in outcome may be due to the quality of therapy provided versus the type of therapy. Assessment of treatment adherence, fidelity, and competence can allow investigators to determine if each therapy is accurately represented (see "Assessment of Treatment Adherence and Integrity" for a more thorough discussion).

Finally, even if the noted threats to internal validity are accounted for in some way, the results of a comparative treatment design can only provide time-limited knowledge. Because each therapy approach and technique is continually being modified and improved upon, outcome can be expected to improve over time. As Borkovec and Miranda (1999) note, this problem applies to any design that solely focuses on applied efficacy "without a commitment or methodology devoted to acquiring basic knowledge through the identification of specific cause-and-effect relationships" (p. 150). Dismantling and constructive study designs, along with studies focusing on treatment mediators and moderators, can better provide the important information regarding mechanisms of change in therapy. Thus, when comparative designs are augmented through the addition of comparison conditions that also allow a dismantling or constructive approach, the researcher can address both immediate public health concerns and questions that are important from a basic science perspective.

Comparisons of Psychotherapy and Medication (and Their Combination)

A specific variation of the comparative treatment design involves the evaluation of the relative efficacy of psychotherapy and medication (or their combination). This design can be attractive because it involves collaboration between researchers from both clinical psychology and psychiatry; many such trials have been conducted in recent years. An interdisciplinary approach to treatment outcome research may encourage more cross-fertilization between the disciplines. However, because this type of collaboration involves researchers with different approaches to treatment, several compromises in study design and treatment protocols are usually involved. In this section of the chapter, we

do not provide answers as much as we raise questions.

Treatment with medication is quite different from treatment with psychotherapy, and these inherent differences raise many unique considerations when comparing these two treatment modalities. For example, when an individual receives medication, if compliance is maintained, he or she receives the treatment on a relatively continuous basis (that is, the active ingredient in the medication is always present in the bloodstream), but in a psychotherapy condition the treatment may be limited to the actual therapy sessions (although homework assignments may extend the effective time per week in therapy). Furthermore, if a research participant misses a psychiatrist visit, but still takes the proper dose of medication, the treatment continues relatively unaffected, but if a participant misses a therapy session, treatment may be stalled or the total amount of treatment reduced. However, it is likely to be impractical to extend the period before posttreatment assessment in order to accommodate makeup sessions since the medication period would need to be similarly extended. The alternative of scheduling multiple therapy sessions per week in order to make up for earlier missed sessions does not allow a sufficient interval between sessions for homework assignments to be effectively executed, a particular problem for action-oriented treatments like CBT. Balancing the different requirements of these very different treatment modalities is a continuous challenge, which only becomes greater when one considers combinations of medication and psychotherapy in addition to simple comparisons.

The nature of medication treatment must also be addressed in the research design. For example, in practice most medications are titrated to an optimal dosage and then maintained at that dosage for a (sometimes extended) period of time. This can be difficult to transfer to a research design that specifies treatment parameters such as length of treatment and dosages in advance. Hollon et al. (1992) conducted a study of various treatments for depression that addressed this issue. The treatments of interest included cognitive therapy (CT), imipramine (a tricyclic antidepressant medication), and their combination during a 12-week treatment period. A fourth condition was also included, in which imipramine was administered for 12 additional months in order to investigate the long-term effects of continuing the medication. After 12 weeks of treatment, CT and imipramine alone produced comparable effects, and combined treatment was not superior to either single modality. However, results from a two-year follow-up suggested that rates of relapse for the 12-week CT group did not differ from the group who received extended medication treatment, which was noted as the current standard practice of prevention, but both groups surpassed the group that received imipramine for only 12 weeks (Evans et al., 1992).

In designs that include discontinuation of medication, gradual tapering is often necessary. Withdrawal effects are common during this period and need to be monitored. Comparable time in the comparison conditions may be included in the design, and this may involve meeting with therapists for additional sessions beyond the number that would typically be included. The content of these sessions, especially in combination treatments, may need to address continued application of the skill set acquired in therapy during the sometimes uncomfortable transition off medication and addressing the patient's attributions for changes that he or she has made in treatment. Assessment will also need to be addressed differently when comparing psychotherapy to a medication treatment. Some assessments may include biological markers of the effects of the medication. Furthermore, what is the best timing of the posttreatment assessment? Is it before or after medication patients have been withdrawn from the drug? These are just some of the many unique considerations involved in comparing psychotherapy to medications.

Treatment-Moderator Design

A treatment-moderator design focuses on what characteristics (beyond the treatment itself) affect the outcome of a given therapy. Moderators of treatment include such diverse factors as patient and/or therapist characteristics or the context in which therapy is provided. In contrast to the goal of a comparative treatment design, which is to investigate whether one therapy is better than another for a particular problem, the goal of a treatment-moderator design is to determine which therapy for a particular problem works best *for whom* or *under what conditions.* Since it is likely that multiple moderators affect outcome (Kazdin & Crowley, 1997),

most of these studies examine several potential moderating factors.

For example, several potential treatment moderators were examined in a study investigating the effects of cognitive therapy for antisocial and aggressive behavior in children (Kazdin & Crowley, 1997). Children's reading achievement, academic and school dysfunction, and number of symptoms across diagnoses affected treatment outcome: children with higher reading levels, who were functioning at a higher level in school, and who had a lower number of symptoms, had a better response to treatment. In addition, socioeconomic disadvantage, parent dysfunction, and aversive parenting practices were contextual factors that moderated outcome.

Treatment-Mediator Design

The treatment-mediator design allows for an investigation of the processes or mechanisms within a treatment package that may contribute to the observed changes. The focus is typically not on the content of the therapy (what is being discussed, what information is presented, and so on), but rather the theorized process of the therapy (the affect, behavior, or cognitions of the client or aspects of the therapist-client relationships).

For example, one study investigated whether changes in cognition and behavioral skills (which are targeted in CBT) mediate the effect of the treatment on depressive symptoms (Kaufman, Rohde, Seeley, Clarke, & Stice, 2005). Specifically, the researchers were interested in testing the theoretical mechanisms that underlie CBT, so they investigated mediators believed to be associated with CBT, such as improved social skills, increased pleasurable activities, the use of relaxation techniques, identification of irrational thoughts and creation of counter-thoughts, and improved problem-solving skills. Changes in thinking appeared to mediate the effects of treatment on depressive symptoms, suggesting that reducing negative thinking may be a primary mechanism by which the intervention produces positive outcomes. Such studies can be useful in evaluating the theoretical basis of the treatment and can also help improve treatment by indicating which factors are most important.

Effectiveness Studies

The study designs described earlier, collectively referred to as "efficacy studies," focus on maintaining internal validity, but this may come at the cost of external validity (that is, the generalizability of the treatment beyond the confines of the research study). *Effectiveness studies* investigate the extent that empirically supported treatments can be effectively applied to everyday practice, or outside the constraining factors of a research study. Factors to be considered in terms of their effect on observed outcome include the characteristics of the study sample, the context in which the therapy is provided, and the novelty of a "new" or innovative treatment.

For example, participants in treatment outcome studies may significantly differ from the individuals who seek treatment in the community. If these groups differ, then the generalizability of a study's findings to general clinical practice is questionable. Study participants may differ from clients in several ways: severity of the presenting problem, presence of other comorbid diagnoses, socioeconomic status, education level, ethnicity, age, and likelihood that they will remain in treatment. These factors can be more or less relevant to specific types of presenting problems. If any of these differences exist, the research sample can be considered biased. Although the bias will not necessarily affect the outcome of treatment, this possibility cannot be ruled out unless it is tested.

In addition to differences in sample characteristics, the context in which the therapy is provided in a research study can differ drastically from the context of therapy provided in the community. For example, participants may be receiving therapy at an academic institution or a specialty clinic, both with a certain reputation that may affect the expectations for treatment. The resources of each context are also likely to differ. Necessary resources for the treatment may be more easily obtained in a research study (e.g., with external funding). However, not all clinics may have these resources, so this may call for a modification of the research study protocol. In addition, therapists in community clinics may also have higher caseloads and may therefore have less time to devote to more involved procedures or preparations (although graduate student or postdoctoral therapists, often employed in research studies, may in some instances have even less time for such preparations due to other commitments).

Therapy in the context of a research study may also be administered differently than the therapy provided in the community. Many

research studies involve the use of a treatment manual (to maintain internal consistency). However, practicing clinicians often do not use manuals, or they may not provide the same type of therapy to all clients with a particular problem. Furthermore, if an empirically supported treatment manual is disseminated for use in the larger clinical community, we cannot assume that practicing clinicians will administer the therapy in the same way as study therapists. Compared to therapists participating in a research study, they may not be as familiar with or invested in the treatment, or they may not believe in the treatment as strongly. Alternatively, they may have a wider range of experiences than study therapists, which may enhance or detract from their administration of the treatment. Many studies investigating the effectiveness of treatments involve extensive training and supervision of community clinic staff. This is necessary if the investigators wish to ensure that the study is testing the administration of the treatment as it has been found to be efficacious. For example, therapist training in an effectiveness study of treatment for depression (Merrill, Tolbert, & Wade, 2003) involved first training four therapists via an intensive program at an expert clinic (the Beck Institute for Cognitive Therapy and Research), followed by those therapists conducting a three-day in-house training workshop, which was then followed by intensive one-on-one supervision. The supervision included feedback from taped sessions and weekly case meetings to ensure treatment integrity. The feasibility of this type of extensive training and supervision outside an effectiveness study is questionable. If researchers are ultimately interested in determining how well treatments can be disseminated to community service clinics outside of research studies, then various degrees and modes of training and supervision also need to be investigated to see how little of this training can be provided without compromising the effects of the treatment.

Several effectiveness studies have used a benchmarking strategy in evaluating the effectiveness of cognitive-behavioral treatments in community settings (Merrill et al., 2003; Stuart, Treat, & Wade, 2000; Wade, Treat, & Stuart, 1998). This strategy involves using the magnitude of change obtained in controlled efficacy studies as the benchmark to which the magnitude of change in community settings is compared. A study investigating CBT for panic disorder administered in a community mental health center (CMHC) used an empirically supported, manualized treatment protocol (Barlow & Craske, 1994), and used the changes reported in previous efficacy studies as the benchmark comparison (Barlow, Craske, Cerny, & Klosko, 1989; Telch et al., 1993). In addition to differences in treatment setting, the effectiveness study differed from the efficacy studies in that no one with a primary diagnosis of panic disorder was excluded from the study on the basis of age, comorbidity, medical problems, treatment history, use of medications, or personality dysfunction. Therapists providing the treatment were CMHC therapists who received extensive training and supervision. The effectiveness of the treatment was demonstrated by rates of improvement among treatment completers similar to those achieved in the efficacy studies, and improvement rates were still comparable at one-year follow-up (Stuart et al., 2000).

In addition to these promising findings, the study revealed differences between the CMHC completers and noncompleters: Education level was significantly lower in the noncompleter group, as compared to both CMHC completers and participants in the original efficacy studies. Despite the additional support provided to these clients (audiotaped versions of the manual, individualized versus group sessions), the less educated clients were less likely to complete treatment. Furthermore, the clients who did not complete treatment were significantly more likely to have one or more comorbid conditions, which contrasts the findings of an efficacy study (Brown, Antony, & Barlow, 1995) that did not find pretreatment comorbidity to predict completer status. Together these results suggest that more research is necessary to determine how to improve treatment for this group of clients. More generally, it is clear that effectiveness research presents significant challenges, yet provides the potential to translate the impressive effects found in efficacy studies to the world of treatment as usual.

Considerations in the Conduct of Treatment Outcome Research

Creation of a Treatment Manual

In scientific endeavors, it is essential that experiments be replicated in order to verify their results. This ideal provides a unique challenge to

treatment outcome research, given that treatment involves a complex interaction between (at least) two people. Practically speaking, it appears impossible to replicate a treatment outcome study unless the treatment is thoroughly defined, usually via a treatment manual.

Although we have certainly heard therapists complain that working from a therapy manual is artificial or constraining, we believe the benefits of creating a manual far outweigh its inconveniences. Without a manual, both trainers and trainees can be left without a clear notion of what the therapy involves. Creating a therapy manual is a valuable way for treatment developers to communicate what the treatment is about, as well as a method for new therapists to come to a similar understanding. Although a difficult process, this specificity need not be a means of robbing a therapy of its lifeblood: at best, it is a means of helping everyone involved come to understand what that lifeblood actually is.

Kendall, Chu, Gifford, Hayes, and Nauta (1998) discuss a number of steps involved in developing manualized treatments that result in a manual that can be used flexibly, is responsive to the individualized needs of clients, and requires clinical skill to be optimally successful. Clinicians must understand the underpinnings of the treatment, not just the specific procedures, so that treatment is guided by theory and not merely techniques. This focus on theory assists clinicians in adjusting the treatment for the individual client. Kendall et al. (1998) assert that there is a middle ground in using treatment manuals between the complete freedom of an unstructured treatment and the strict adherence to every detail of a rigidly structured manual. Accordingly, we suggest that in creating a manual, specificity is necessary to allow others to provide the therapy, but excessive specificity can be a barrier to learning and encourages the rote repetition of nonefficacious aspects of treatment. Therapist manuals are at their best when they are specific about the aspects of therapy that are essential, are clear about what therapeutic techniques are and are not involved in the treatment, and are accessible to therapists who have little prior experience with the type of therapy (or provide guidelines as to how otherwise professionally qualified novices can prepare themselves). Especially as therapy manuals are revised, it is useful to expand upon typical pitfalls and difficulties that therapists encounter and how they can be avoided or resolved.

As mentioned previously, we have often encountered therapists who believe that therapy manuals obstruct treatment. Although an incompetently rendered manual or a manual for an ineffective treatment could certainly do so, we do not believe that well-written manuals for effective treatments have this effect. Such manuals will certainly restrain therapists, to greater or lesser degrees, from doing what they would do if they had not encountered the manual. Far from being artificial or limiting, however, such constraint is essential to the delivery of an effective therapy. Without a manual, therapists may behave in ways that they would consider innocuous (or not consider at all), whereas these behaviors may actually have an effect (for better or for worse). The creation of a treatment manual is one step toward ensuring that only the ingredients specified in the therapy are delivered: Therapist training and evaluation of treatment integrity are two more.

Therapist Training and Supervision

The creation of a treatment manual is a first step toward providing training to therapists. As with the creation of a treatment manual, there is no single therapist training procedure that is sufficient for every context. We would suggest, at a minimum, that trainees have adequate professional training, interact with trained therapists in a structured way (e.g., through training sessions), demonstrate initial competence with therapy techniques via role play, and treat some number of clients with the therapy protocol while communicating with experienced supervisors on a regular basis. The last step will best accomplished by providing supervisors with audio- or videotapes of therapy sessions in advance of supervision meetings.

Assessment of Treatment Adherence and Integrity

When testing whether a therapy or a specific component of a therapy has a particular effect, it is essential to be sure that the therapy or therapy component has been delivered in the active condition and not delivered in the inactive condition (*adherence*) and that the therapy has been delivered competently or as intended by the investigators (*integrity*). A standard practice for providing this assurance involves audio- or

videotaping all therapy interactions and selecting, at random, a subset of these tapes for rating of treatment adherence and integrity. Many studies (e.g., Barkley, Edwards, Laneri, Fletcher, & Metevia, 2001; McDonagh et al., 2005) have rated about 25% of therapy interactions. Although more complete rating provides more complete assurance, the primary concern for an investigator who wishes to demonstrate treatment integrity is to be sure that a reasonably large sample of interactions is rated.

Several issues arise regarding treatment adherence and integrity ratings. First, ratings require a clear operationalization of what is and is not a part of the treatment condition. An appropriately specific treatment manual will make this process easier by clearly describing these components, but rating scales must be more specific and focus on what can be seen and/or heard by raters. If, as is sometimes the case, the raters are not therapists experienced in the given treatment, the rating scales must be particularly clear. Second, as with all types of ratings, their usefulness is uncertain unless evidence of their reliability is supplied (e.g., by having two or more raters rate the same sessions). Third, when feasible, such ratings can be used to verify the adherence and competence of current therapists, providing a mechanism for improving treatment adherence and integrity or eliminating therapists who are unable to provide adequate treatment.

Depending on the design of the study, treatment adherence may have different meanings. If only one treatment is being provided, evaluation of adherence involves testing that the therapists provided what was specified as necessary and avoided what was specified as undesirable. If two (or more) treatments are being compared, particular attention can be given to determining that one treatment has not been inadvertently contaminated with elements that are believed to be unique to the other therapy. A particularly conservative approach to such ratings would be to provide raters with a randomized list of therapist behaviors, some of which belong in one treatment, some in another, and potentially some that belong to both and/or neither (see Hill, O'Grady, & Elkin, 1992). Raters who are blind to what treatment is supposed to be present can then use such a rating scale to provide feedback that is uncontaminated by expectations. If raters are also not trained in the therapies involved, such a system of rating treatment adherence may be capable of bypassing the assumptions clinicians sometimes have about what elements belong together in a treatment. A recent study (Ablon & Jones, 2002) used a similar methodology, but used a Q-set instrument that consisted of 100 items referring to therapist-patient interactions. The items were sorted into a continuum from least characteristic to most characteristic of a particular therapy session. Session ratings were compared to prototypes of ideal treatments—CBT and IPT—that were constructed by panels of expert therapists in the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH TDCRP). Ratings indicated that both treatment conditions were more similar to the CBT prototype. Furthermore, adherence to the CBT prototype correlated with positive treatment outcome in a more consistent and robust manner for both the IPT and CBT conditions. This study draws attention to the importance of assessing treatment adherence in testing the assumption that any improvements in treatment studies are due to the techniques described in the treatment manuals.

As noted previously, treatment integrity includes therapist competence: the skillfulness of the therapist in providing the treatment, conceptualizing the patient's problems within the treatment framework, and applying techniques and methods consistent with the treatment rationale (Shaw et al., 1999). Expert judges who review the work of the therapists typically make these ratings. For example, in a study investigating the relationship between therapist competence and outcome of cognitive therapy for depression (Shaw et al., 1999), therapy sessions were videotaped and then sent to three experts who trained the therapists in CBT for the NIMH TDCRP. The raters then rated the sessions using a scale that measures cognitive-behavioral therapist competence (Cognitive Therapy Scale; Young & Beck, 1980). The scale produced two factors: skill (e.g., understanding, interpersonal effectiveness, collaboration, focusing on cognitions) and structure (e.g., setting an agenda, session pacing, homework assignment and review). Results provided some limited support for the role of therapist competence (particularly structure) in reducing depressive symptoms, but this relationship was only significant for a clinician-rated measure of symptoms, not self-reported symptoms. The authors discussed how the therapist

competence rating scale may not have been sufficiently comprehensive and suggest that "the challenge for the field is to improve the measures of therapist skillfulness to be used in future research, so that we can more clearly identify the actual components of therapist behavior that lead to favorable outcome" (p. 845).

Reduction of Missing Data

Most psychological treatments involve multiple sessions, and even the simplest treatment study involves pre- and posttreatment measurement. At each assessment point, it is possible to lose data. Although some missing data in large-scale studies is inevitable, treatment outcome studies present particular challenges to the researcher who is interested in obtaining the most complete data possible.

Typically, participants are initially compliant with requests for data, partially because they expect they will receive a possibly beneficial treatment and do not wish to alienate the professionals who might provide it to them. Although this is at first helpful to the researcher, it may soon backfire when participants either gain all the benefit they believe they need or start to believe that they will not derive this degree of benefit. It behooves the careful researcher to account for these possible changes in motivation by providing additional incentives for participants to complete assessments. Where possible, a small amount of financial compensation may help reduce missing data. Employing staff that monitor participants' status in the study and encouraging them to attend assessments can also be useful.

Multimodal/Informant Assessment

After the completion of treatment, outcome is assessed. Often the outcome measure of greatest interest assesses the presenting problem that first brought the individual to treatment. However, most presenting problems are multifaceted, and multiple modes of assessment will provide investigators with a more accurate measurement of the outcome of interest. An individual can provide his or her own report of symptoms. In addition, an interviewer can obtain information about symptomatology and use clinical judgment to determine which symptoms remain problematic. Because presenting problems often involve a behavioral component, behavioral observation can also be used. For example, communication with a marital partner, parenting behaviors, tics, public speaking, and in-school oppositional behavior can all be observed and assessed both before and after treatment. Psychophysiological measures are another mode of assessment (see Chapter 11 of this volume) that has become easier for researchers to use due to advances in the relevant technology. Heart rate, blood pressure, and skin conductance can be assessed to gain a direct measure of physiological responses. Neuroimaging can also provide valuable information about neurological processes involved in psychopathology and changes in this arena that come with successful treatment.

Obtaining information from various sources can also provide a more accurate assessment of outcome. This may involve gathering information from a participant's family member or spouse or a child's teacher or parent. Clinical interviewers, mentioned earlier, are frequently used in treatment studies as a primary source of information. To reduce any bias in evaluation that may occur due to knowledge about the participant's treatment condition, assessors may be kept blind to this information. An independent assessor can interview participants about their progress and make the evaluation of change based on the provided information and observations while minimizing the influence of their beliefs about the effectiveness or appropriateness of a given treatment condition. The use of independent assessors adds complexity to the study design: therapists cannot serve as evaluators, and both assessors and participants must make efforts to maintain the blind. However, otherwise, the validity of clinician assessments is unclear: Is the interviewer responding to the client or some combination of the client and the treatment condition information they already have about the client? Similarly, if the treating clinician provides the ratings, can he or she do so in a manner that is independent of his or her biases in favor of that treatment?

Assessment Throughout Treatment

Much of the discussion thus far focused on measuring posttreatment functioning and comparing outcome across treatment conditions. However, assessing progress throughout treatment adds substantially to a treatment outcome study. If midtreatment assessments are not

obtained, the evaluation of the treatment is limited in that the results apply only to changes seen when the treatment is complete. However, there may be meaningful changes that occur throughout the therapy process that are undocumented when assessment is limited in this way. The earlier discussion of treatment mediators addressed the benefits of determining mechanisms of change or how the therapy produces a beneficial outcome. Assessments over the course of treatment can provide some of this important information about theorized treatment mediators. For example, observing improvements in both social anxiety and depression at the end of treatment does not provide any information about the relationship between these two observed changes. In contrast, if both social anxiety and depression are assessed throughout treatment, conclusions can be drawn about which change may have led to the other change. A recent study reported on the specific process of changes in social anxiety and depression during treatment for social anxiety disorder (Moscovitch, Hofmann, Suvak, & In-Albon, 2005). Symptoms were assessed weekly during the course of a 12-week group treatment program. Decreases in social anxiety fully mediated decreases in depression (accounting for 91% of improvements), while decreases in depression only partially mediated decreases in social anxiety (accounting for 6% of improvements). In other words, during treatment for social anxiety disorder, depression improves because social anxiety improves.

Essentially, the course of a treatment study can be treated as any other time period in a longitudinal design. Longitudinal researchers have developed powerful methods for examining change in and causal relationships among variables (e.g., growth curve modeling, autoregressive models, or their synthesis, the autoregressive latent trajectory model; Curran & Bollen, 2001). However, these methods can only be used effectively when multiple assessments are available; in general, more assessments are preferable, so long as the number of assessments does not burden participants in an unethical or treatment-interfering manner.

Assessment After Completion of Treatment

Information about the long-term benefits of treatment is crucial in evaluating any treatment package. A treatment that produces large effects by the end of treatment is not as impressive if the effects are transitory. Follow-up assessments allow tests of long-term maintenance of improvement. It is usually best to conduct a complete assessment (similar to the assessment conducted post-treatment) at follow-up increments that are suitable to the disorder being studied. Obtaining information about any other treatment received during the follow-up assessment is also helpful, given that additional treatment is a potential confounding variable. If all participants maintain their gains, but nearly all receive additional treatment, the effects may no longer be attributable to the study treatment alone. For example, Heimberg, Salzman, Holt, and Blendell (1993) noted the lack of adequate follow-up assessments in studies of treatment for social anxiety disorder. They reported on five-year follow-up data for individuals who received either cognitive-behavioral group therapy or educational supportive group therapy for social anxiety disorder. Participants were evaluated with self-report questionnaires, structured clinical interview, and a behavioral test. Five years after treatment, participants who received CBGT remained more improved than those who received ES on measures from all assessment modalities.

Obtaining crucial follow-up information can be difficult. The largest difficulty is the loss of participants willing to participate in the study after active treatment has ended. For example, in the study just discussed, of 37 study participants, 5 could not be contacted and 13 declined participation. It may be difficult or impossible to obtain information from all original participants for a variety of reasons, and the inability to obtain information from all participants may bias the follow-up results. The individuals who do not return for follow-up assessment may differ from the follow-up completers in meaningful ways, such as symptom severity or rates of relapse. Potential differences between follow-up participants and nonparticipants were examined by Heimberg, Salzman, Holt, and Blendell (1993). Most analyses yielded insignificant differences, but some differences did emerge suggesting that compared to the nonparticipants, participants who completed five-year follow-up may have been less severely impaired before treatment and at the six-month follow-up assessment. Therefore, the authors indicate that CBGT appears to be effective in the long-term for participants, but these results may be limited to those who were initially less severely impaired.

Assessing Clinically Significant Change

A central question in treatment outcome research is "Does a particular therapy produce meaningful changes?" However, what are *meaningful changes*? Posttreatment symptom levels may be statistically different across different treatment conditions, but the presence of any particular change in symptoms does not guarantee that participants have experienced meaningful changes in their lives.

One avenue toward answering this question involves asking participants (or people close to the participants) to report on how important the changes (if present) are, or how much of a difference (if any) the treatment has made in daily life. The importance of subjective satisfaction with a treatment is intuitively appealing; however, there are some problems with this approach. For example, perceived changes do not necessarily mean that the presenting problem has been effectively treated. Also, reports of improvement or satisfaction may be particularly biased: any given individual has likely invested a considerable amount of time in the treatment and, due to this investment, may be prone to endorse the treatment as effective. Similarly, participants who have developed a strong alliance with a therapist may feel reluctant to admit that a treatment has failed to result in meaningful change.

Reports need not be limited to reports of satisfaction and changes in symptoms. Assessing an individual's quality of life is way to broaden evaluation of treatment outcome and clinical significance (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999). The assessment of quality of life is clouded by the lack of a clear, consensual concept. One predominant view (see Diener, 1984, for a review) equates quality of life with one's satisfaction with various life domains that are personally important. As the definition implies, quality of life goes beyond the presence or absence of presenting symptoms. Assessment of quality of life sheds more light on the clinical significance of changes, but it is typically assessed via self-report, so it is subject to the same biases discussed previously. In addition, ratings of quality of life appear to be dependent on affective state. Correlations between satisfaction and mood ratings raise concerns about the possible redundancy of quality of life ratings and symptom measures (Gladis et al., 1999).

Another approach to determining clinical significance involves comparing individual outcome to an objective standard, such as mean scores on key measures in normative samples (see Chapter 13 of this volume). A clinically significant change would be likely if a treatment produced change that returned the participant to a normative level of functioning (assuming that pretreatment level of functioning differed from the norm). As Kendall, Marrs-Garcia, Nath, and Sheldrick (1999) point out, normative comparisons can provide a verification of the meaningfulness of the difference between treatment groups, and when comparing two treatments that are both effective according to statistical tests, can potentially determine which of the treatments is preferable. However, defining normative behavior can be difficult and may not always be a meaningful comparison. Because psychiatric disorders and symptoms are present in the general community and "normal" samples, the implications of posttreatment functioning within the normal range can be unclear, depending upon the problem of interest.

Alternatively, a treatment group could be compared to some type of clinical or dysfunctional sample. Examples include comparing each individual's posttreatment scores to mean scores of an untreated control group or to pretreatment means. If a clinically significant change were achieved, then the individual would differ greatly from a nontreated or dysfunctional sample. Various criteria are used within the field to determine clinical significance. For example, Jacobson, Roberts, Berns, and McGlinchey (1999) have suggested using two standard deviations from the mean of the dysfunctional sample as one stringent criterion for defining clinically significant change: if an individual scores two standard deviations above the mean of an untreated control group, he or she is no longer represented by that mean and distribution.

If a treatment targets a particular disorder, clinical significance can be assessed by determining if an individual continues to meet diagnostic criteria. If participants are recruited on the basis of a particular disorder, then we can assume 100% of them met criteria for that disorder before treatment. Rates of disorder after treatment can be compared across groups. Even with this approach, it is wise to not assume that failing to meet diagnostic criteria represents clinically significant change. Reduction in a single symptom may preclude the diagnosis, but an individual may still be impaired by remaining symptoms. For example, finding that a participant no longer meets

criteria for major depressive disorder seems less than impressive if the individual currently meets criteria for dysthymic disorder.

Finally, clinical significance can be determined by assessing the degree of social impact the treatment has made on outcomes that are important to society at large. Examples include rates of hospitalization, illness, suicide, criminal behavior, and employment. Documented changes on social impact variables are usually of particular interest to policymakers because these variables are more applicable to society as a whole. Such indices can be more easily interpretable to the public at large.

Attrition and Intent-to-Treat Analyses

Attrition, also referred to as *dropout,* refers to the loss of participants over the course of the study. Attrition can occur at various stages of treatment or follow-up, and participants may decide to discontinue treatment for various reasons. Random assignment to treatment conditions can be assumed to distribute reasons for dropout equally across groups if samples are of sufficient size. However, attrition rates are usually compared across groups to ensure that one condition does not produce more dropouts due to the nature or effects of the treatment. Treatment completers can also be compared to those who discontinued treatment on pretreatment and midtreatment measures. Such a comparison can illuminate factors that may contribute to dropout. However, unanticipated and therefore unmeasured variables may have contributed to dropout, and these factors may confound the interpretation of differential treatment effects.

Because some participants will discontinue treatment and hence not receive the complete treatment package, the selection of appropriate data to include in subsequent analyses is complex. In the psychological treatment literature, the tradition has been to focus upon those who complete treatment (or an equivalent amount of time in the case of no-treatment or wait-list control conditions). The rationale for this approach is that the best way to evaluate whether a treatment is effective is to examine the impact it has when delivered at maximum dose. In the psychiatric treatment literature, the default position has been the *intent-to-treat analysis.* This approach includes all persons assigned to a treatment condition, regardless of the amount of treatment they received. This is typically accomplished by using the last observation carried forward, which involves repeating the last collected score at subsequent assessment points, and thus assuming no further change. This approach tells us less about the efficacy of the treatment at maximum dose, but it may tell us more about the utility of the treatment in the larger population. It may also provide more information about the impact of treatment from a public health perspective. However, it is still unclear how similar outcomes are when examined in treatment completers versus the intent-to-treat approach, or how missing data and early attrition affects the validity of analyses using either approach. There may be benefits to presenting both types of analyses in examining outcomes.

Statistical Concerns: Use of Multivariate Techniques and Analysis of Covariance

Although it is the last topic we are considering, we suggest that statistical concerns should be considered far in advance of conducting treatment outcome research. Identification of particular statistical techniques that can supply the best fit to the underlying theoretical model will often result in greater specificity as to how data should be collected. If, as we recommend, multiple measures of outcome from multiple sources are obtained, multivariate analyses of variance (MANOVA) or other multivariate techniques are the preferred statistical tests. Treatment outcome research has often used univariate analysis of variance (ANOVA), given that this research often employs multiple randomly assigned conditions. Evaluation of many individual measures compared across treatment conditions multiplies the possibility of statistically significant differences due to chance. Using a MANOVA approach, analyses are conducted with a combination of various measures of the same construct. If a significant effect is found in the MANOVA, follow-up ANOVAs can be conducted to determine which specific measures differ across conditions.

An attractive alternative to combining measures according to a priori assumptions about the proposed constructs they assess involves conducting a factor analysis of all measures and combining the measures that form a factor. If a priori assumptions are available, confirmatory factor analysis can be used in order to assess how well the proposed structure fits the data. If no a priori assumptions exist regarding the particular

methods used, exploratory factor analysis may be used to determine how to best combine measures. Finally, in designs that involve continuous predictors and multiple outcome variables, structural equation modeling approaches can allow multivariate tests of the relationships between predictors and the latent factors derived from multiple outcome variables, bypassing the need for extensive univariate testing and providing statistical tests closely tied to the underlying theoretical model.

One further statistical concern is the misuse of analysis of covariance (ANCOVA) to address or control for group differences on potential covariates (Miller & Chapman, 2001). When groups do not differ on covariates, the use of ANCOVA can effectively serve to reduce the error variance (or noise in measurement), and it is a legitimate but underutilized statistical tool. On the other hand, ANCOVA is more commonly used when groups differ on covariates in order to control for the differences; however, this is an inappropriate use of ANCOVA. Because the grouping variable or treatment condition is correlated with the covariate, it is unclear what remains when the effect of the covariate is removed; the variable is altered in a substantial way that is not specifiable. Miller and Chapman provide an example. Attempts to separate, or covary, the effects of depression and anxiety are complex because they share symptoms and psychological and neural processes. If two groups (one depressed, one control) differ on measures of anxiety, treating anxiety as a covariate when comparing the groups will *not* result in a depressed group that is a representation of depression as it exists without anxiety. Overall, we encourage researchers to carefully consider how their research hypotheses are best specified via statistical tests: it seems unlikely that a researcher would really be interested in depression that is completely unassociated with anxiety, for example, because such a construct does not seem to occur in the real world.

Conclusion

In summary, there are many design, assessment, and statistical considerations in treatment outcome research. Knowledge of how a treatment study addresses these factors is essential in evaluating its findings. Consideration of the factors addressed in this chapter also helps an investigator design the best treatment study to evaluate a new or already developed treatment program. Any treatment outcome study will involve significant amounts of time and resources on the part of the researcher, as well as an investment of time and hope on the part of participants. Therefore, careful consideration to design is necessary to develop a study that will ultimately provide answers to questions of interest, make the best use of allocated resources, and treat participants in an equitable and ethical manner.

References

Ablon, J. S., & Jones, E. E. (2002). Validity of controlled clinical trials of psychotherapy: Findings from the NIMH Treatment of Depression Collaborative Research Program. *American Journal of Psychiatry, 159,* 775–778.

Barkley, J., Edwards, G., Laneri, M., Fletcher, K., & Metevia, L. (2001). The efficacy of problem-solving communication training alone, behavior management training alone, and their combination for parent-adolescent conflict in teenagers with ADD and ODD. *Journal of Consulting and Clinical Psychology, 69,* 926–941.

Barlow, D. H., & Craske, M. G. (1994). *Mastery of your anxiety and panic II.* Albany, NY: Graywind.

Barlow, D. H., Craske, M. G., Cerny, J. A., & Klosko, J. S. (1989). Behavioral treatment of panic disorder. *Behavior Therapy, 20,* 261–282.

Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. K. (2000). Cognitive-behavior therapy, imipramine, or their combination for panic disorder: A randomized-control trial. *Journal of the American Medical Association, 283,* 2529–2536.

Barrett, P. M., Dadds, M. R., & Rapee, R. M. (1996). Family treatment of childhood anxiety: A controlled trial. *Journal of Consulting and Clinical Psychology, 64,* 333–342.

Borkovec, T. D. (1994). Between-group therapy outcome research: Design and methodology. In L. S. Onken & J. D. Blaine (Eds.), *NIDA Research Monograph #137* (pp. 249–289). Rockville, MD: National Institute of Drug Abuse.

Borkovec, T. D., & Miranda, J. (1999). Between-group psychotherapy outcome research and basic science. *Journal of Clinical Psychology, 55,* 147–158.

Brown, T. A., Antony, M. M., & Barlow, D. H. (1995). Diagnostic comorbidity in panic disorder: Effect on treatment outcome and course of

comorbid diagnoses following treatment. *Journal of Consulting and Clinical Psychology, 63,* 408–418.

Budney, A. J., Higgins, S. T., Radonovich, K. J., & Novy, P. L. (2000). Adding voucher-based incentives to coping skills and motivational enhancement improves outcomes during treatment for marijuana dependence. *Journal of Consulting and Clinical Psychology, 68,* 1051–1061.

Butler, G., Cullington, A., Munby, M., Amies, P., & Gelder, M. (1984). Exposure and anxiety management in the treatment of social phobia. *Journal of Consulting and Clinical Psychology, 52,* 642–650.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 105–136). Washington, DC: American Psychological Association.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95,* 542–575.

Evans, M. D., Hollon, S. D., DeRubeis, R. J., Piasecki, J. M., Grove, W. M., Garvey, M. J., et al. (1992). Differential relapse following cognitive therapy and pharmacotherapy for depression. *Archives of General Psychiatry, 49,* 802–808.

Foa, E. B., Franklin, M. E., & Moser, J. (2002). Context in the clinic: How well do cognitive-behavioral therapies and medications work in combination? *Biological Psychiatry, 52,* 989–997.

Gladis, M. M., Gosch, E. A., Dishuk, N. M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 320–331.

Glass, C. R., Gottman, J. M., & Shmurak, S. H. (1976). Response-acquisition and cognitive self-statement modification approaches to dating-skills training. *Journal of Counseling Psychology, 23,* 520–526.

Heimberg, R. G., Dodge, C. S., Hope, D. A., Kennedy, C. R., Zollo, L. J., & Becker, R. E. (1990). Cognitive-behavioral group treatment of social phobia: Comparison to a credible placebo control. *Cognitive Therapy and Research, 14,* 1–23.

Heimberg, R. G., Liebowitz, M. R., Hope, D. A., Schneier, F. R., Holt, C. S., Welkowitz, L., et al. (1998). Cognitive-behavioral group therapy versus phenelzine in social phobia: 12-week outcome. *Archives of General Psychiatry, 55,* 1133–1141.

Heimberg, R. G., Salzman, D. G., Holt, C. S., & Blendell, K. A. (1993). Cognitive-behavioral group treatment for social phobia: Effectiveness at five-year followup. *Cognitive Therapy and Research, 17,* 325–339.

Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavioral therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology, 60,* 73–79.

Hollon, S. D., DeRubeis, R. J., Evans, M. D., Wiemer, M. J., Garvey, M. J., Grove, W. M., et al. (1992). Cognitive therapy and pharmacotherapy for depression: Singly and in combination. *Archives of General Psychiatry, 49,* 774–781.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300–307.

Katz, L. Y., Cox, B. J., Gunasekara, S., & Miller, A. L. (2004). Feasibility of Dialectical-Behavior Therapy for suicidal adolescent inpatients. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 276–282.

Kaufman, N. K., Rohde, P., Seeley, J. R., Clarke, G. N., & Stice, E. (2005). Potential mediators of cognitive-behavioral therapy for adolescents with comorbid major depression and conduct disorder. *Journal of Consulting and Clinical Psychology, 73,* 38–46.

Kazdin, A. E. (2001). Progression of therapy research and clinical application of treatment require better understanding of the change process. *Clinical Psychology: Science and Practice, 8,* 143–151.

Kazdin, A. E. (2004). *Research design in clinical psychology.* Boston: Allyn & Bacon.

Kazdin, A. E., & Crowley, M. J. (1997). Moderators of treatment in cognitively based treatment of antisocial children. *Cognitive Therapy and Research, 21,* 185–207.

Kendall, P. C., Chu, B., Gifford, A., Hayes, C., & Nauta, M. (1998). Breathing life into a manual. *Cognitive and Behavioral Practice, 5,* 177–198.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 285–299.

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

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology, 71,* 404–409.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48.

Moscovitch, D. A., Hofmann, S. G., Suvak, M. K., & In-Ablon, T. (2005). Mediation of changes in anxiety and depression during treatment of social phobia. *Journal of Consulting and Clinical Psychology 73,* 945–952.

Posternak, M. A., & Miller, I. (2001). Untreated short-term course of major depression: A meta-analysis of outcomes from studies using wait-list control groups. *Journal of Affective Disorders, 66,* 139–146.

Rodebaugh, T. L. (2004). I might look OK, but I'm still doubtful, anxious, and avoidant: The mixed effects of enhanced video feedback on social anxiety symptoms. *Behaviour Research and Therapy, 42,* 1435–1451.

Schmidt, N. B., Woolaway-Bickel, K., Trakowski, J., Santiago, H., Storey, J., Koselka, M., et al. (2002). Dismantling cognitive-behavioral treatment for panic disorder: Questioning the utility of breathing retraining. *Journal of Consulting and Clinical Psychology, 68,* 417–424.

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmstead, M., Vallis, M. T., Dobson, K. S., et al. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 67,* 837–846.

Spiritu, A., Stanton, C., Donaldson, D., & Boergers, J. (2002). Treatment-as-usual for adolescent suicide attempters: Implications for the choice of comparison groups in psychotherapy research. *Journal of Clinical Child and Adolescent Psychiatry, 31,* 41–47.

Stuart, G. L., Treat, T. A., & Wade, W. A. (2000). Effectiveness of an empirically based treatment for panic disorder delivered in a service clinic setting: 1-year follow-up. *Journal of Consulting and Clinical Psychology, 68,* 506–512.

Telch, M. J., Lucas, J. A., Schmidt, N. B., Hanna, H. H., Jaimez, T. L., & Lucas, R. A. (1993). Group cognitive-behavioral treatment of panic disorder. *Behaviour Research and Therapy, 31,* 279–287.

Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology, 66,* 231–239.

Young, J., & Beck, A. T. (1980). *Cognitive Therapy Scale rating manual.* Unpublished manuscript. University of Pennsylvania, Philadelphia.