TREATMENT VALIDITY FOR INTERVENTION STUDIES

Dianne L. Chambless and Steven D. Hollon

That psychologists conduct sound research is important whatever their research areas, but it is especially important when the results of this research have a substantial impact on people's lives. This is the case for intervention research, the findings of which may influence the type of treatments people receive. Sound intervention research is critical for determining what treatments are beneficial and what treatments are ineffective or even harmful. Moreover, intervention research is typically very expensive to conduct, with a single project often taking 5 to 7 years to complete and costing well over \$1 million. This means that relatively little psychosocial intervention research is conducted, making the validity of each trial that much more important.

Intervention research is challenging, and during such a study (often called a *trial*), investigators will be faced with many uncontrollable events they could not have foreseen. Common problems in treatment research are predictable, however, and many potential difficulties in interpretation of findings can be avoided with proper attention to design before the trial begins. In this chapter we will review the most common design questions in intervention research and will specify the basic elements we believe must be present to permit researchers to draw valid conclusions from their data.

ASSESSMENT

Assessment is something of a stepchild for intervention researchers. Researchers often pay inadequate

attention to whether the assessment measures they select reliably and validly assess the constructs they wish to measure. There is an unfortunate tendency to choose measures simply because they are in widespread use or because the names of the measures suggest that they represent the constructs of interest. Without valid and appropriate measurement, the results of a study are severely compromised if not meaningless, and time spent up front on selection of appropriate measures will pay off in the end. This is not a task to pass off to an inexperienced research assistant.

Several types of assessment are important: First, how will the investigator determine that the participants represent the types of people to whom this research is supposed to generalize (a critical aspect of the external validity of the study)? In prevention research, the potential participant pool may be everyone who was exposed to some stressor such as a hurricane or all couples attending a given church who are engaged to be married. In such cases, the definition of the sampling frame may be relatively straightforward, although in cases of exposure to stressors, the researcher will want to carefully assess the degree of exposure. In psychotherapy research, it is common, although not mandatory, for the participant pool to be defined in terms of the primary diagnosis conferred, with predefined exclusions of those who have other conditions that might render the proposed treatment inappropriate. For example, investigators typically exclude patients with a history of psychosis from trials of

Preparation of this manuscript was supported in part by NIMH Grant No. K02 MH01697 awarded to Steven D. Hollon.

DOI: 10.1037/XXXXX.XXX

APA Handbook of Research Methods in Psychology: Vol. 2. Quantitative, Qualitative, Neuropsychological, and Biological, H. Cooper (Editor-in-Chief)
Copyright © 2012 by the American Psychological Association. All rights reserved.

nonpsychotic disorders. In such research, the researcher needs to convince readers that the diagnoses were reliably and validly made. For studies of psychiatric disorders, this generally means that structured diagnostic interviews were used and that the researcher demonstrated satisfactory inter-rater reliability for diagnoses by having a randomly selected sample of diagnostic interviews repeated or rated by a second diagnostician who is not informed of the first diagnostician's decisions. Alternatively, participants may be selected on the basis of their scores on self-report measures. For example, couples may be selected for marital therapy on the basis of their stated desire for couples therapy and their falling below some threshold of marital satisfaction on a psychometrically sound inventory. In a research report, investigators are encouraged to clearly report the reasons that potential participants were excluded from the study, decided against participation, or later dropped out, along with numbers for those categories of people, perhaps using a chart advocated by the CONSORT guidelines to standardize the report (Altman et al., 2001). Figure 28.1, taken from Striegel-Moore et al.'s (2010) study on guided self-help for recurrent binge eating, provides an example of a CONSORT chart.

Second, how will the investigator know whether the intervention had the desired effect? Here is it crucial to select reliable and valid outcome measures appropriate to the sample being studied. Because each method of measurement captures only a piece of the latent variable the researcher wants to assess, it is preferable to use multiple methods of assessment, including, for example, self-report, interviewer, and observational or behavioral measures. In research on children, it is desirable to obtain ratings from parents and teachers as well as from the children themselves (Kazdin, 2003; Kraemer et al., 2003). When researchers use interviewer and observational measures, they must demonstrate adequate inter-rater reliability for all occasions of assessment (e.g., pretest, post-test, follow-up). If, as is usually the case, it is too expensive to have reliability ratings for every participant assessed, investigators should randomly sample from all occasions of assessment for reliability ratings and report these findings using

measures of reliability appropriate for the level of measurement (e.g., Cohen, 1960; Shrout & Fleiss, 1979).

Third, how will the investigators know whether the intervention works for the reasons that they propose? Reliable and valid measurement is required to test such process or mediational questions (Baron & Kenny, 1986). For example, if the researcher proposes that psychotherapy works because the client forms a close working alliance with the therapist (e.g., Bordin, 1979), a good measure of the working alliance must be included.

Careful determination of all of the constructs that need to be assessed in the study and selection of appropriate means of measurement are critical to the ultimate success of the trial. Attention to the validity of assessment speaks to the construct validity of the research, a topic we will develop further in a later section.

SELECTION OF THE RESEARCH DESIGN

Randomization

Single-case and quasi-experimental designs will be covered in other chapters. Here we concentrate on the randomized controlled trial (RCT), which permits the strongest inferences of causality, that is, allows the investigator to say with the most confidence that any changes observed are due to the intervention (Shadish, Cook, & Campbell, 2002). In such a design, each individual participant is assigned at random to an intervention condition; alternatively, randomization occurs at some group level. For example, classrooms of children may be assigned at random to a prevention program or alternative condition, or wards of a hospital may be assigned at random to an experimental procedure, whereas other wards serve as a waiting list control condition. This approach to assignment is sometimes called *cluster randomized assignment* (see Campbell, Elbourne, & Altman, 2004, for a discussion of methodological issues in such studies). Randomization is the best method for guarding against *selection* effects (Shadish et al., 2002), the presence of systematic differences between groups that imperil the investigator's ability to draw the conclusion that the intervention rather than preexisting

FIGURE 28.1. An example of a CONSORT flowchart. From “Cognitive Behavioral Guided Self-Help for the Treatment of Recurrent Binge Eating,” by R. H. Striegel-Moore, G. T. Wilson, L. DeBar, N. Perrin, F. Lynch, F. Rosselli, and H. C. Kraemer, 2010, *Journal of Consulting and Clinical Psychology*, 78, p. 314. Copyright 2010 by the American Psychological Association.

differences between groups led to the observed difference in outcome.

When very large numbers of participants are involved, the experimenter can be relatively

confident that randomization will ensure that treatment conditions will not differ on important variables other than receipt of the intended intervention. However, intervention trials are often not so large,

and fairly frequently treatment groups will be found to differ significantly on one or more variables before treatment. For this reason, investigators would do well to consider in advance the presence of other variables that might be related to outcome and on which participants might differ. For example, suppose the investigator believes that patients with borderline personality disorder are likely to do worse in treatment for major depressive disorder than patients without such personality pathology. In this case the investigator might block or stratify the patients on presence or absence of borderline personality disorder before randomization and then conduct randomization within blocks. Such a practice makes it more likely that each treatment condition will have roughly equal numbers of patients with this personality disorder, particularly if the researcher uses procedures to foster balance across conditions or blocks within conditions, among the most common of which is urn randomization (Wei & Lachin, 1988). In urn randomization, the probability that subsequent patients will be assigned to a specific treatment condition is adjusted on an ongoing basis to reflect the number and type of patients previously assigned to that condition. Urn randomization is used to decrease the likelihood that randomization will fail (distribute uneven numbers or types of patients across the treatment conditions) simply on the basis of chance. Knowledge of the literature on factors associated with treatment outcome will guide the investigator in the selection of the most important blocking variables, in that it is not possible to stratify on numerous factors.

In implementing randomization the researcher needs to separate knowledge of eligibility from knowledge of the randomization sequence. Randomization should occur only after it has been determined that a patient meets eligibility criteria. Otherwise the investigator risks having unintentional biases creep in. For example, suppose the investigator is handling random assignment and knows that the next patient in the randomization sequence will be assigned to his or her favored treatment. When that patient proves to be a difficult case the investigator suspects will not do well, he or she might unintentionally find a reason this patient is not eligible for the study. Procedures should be in

place to prevent such temptation, for example, by having someone not involved in eligibility determination (e.g., the project data manager) maintain the randomization sequence in secret, providing the patient's treatment condition assignment only once a firm decision has been made regarding eligibility, and the patient's status on any blocking variables is known (Efron, 1971).

The Comparison Group

Selection of the appropriate comparison group should be determined by the research question. Do we simply want to know whether a treatment is beneficial? If so, the control condition for the RCT can be a waiting list control group that accounts for potential confounding variables, such as the effects of the assessment procedures and the passage of time, during which the problem might run its course or be affected by healing agents in the patient's natural environment. Such a design tells us whether our treatment is *efficacious*, that is, better than no treatment. We might be dissatisfied with this information and want to know whether our treatment works better than some placebo or basic treatment. If it does, then we might call the treatment *efficacious and specific* (Chambless & Hollon, 1998). The design testing specificity allows the investigator to control for additional variables, such as hope and expectancy of change, a caring relationship with a professional, and education about one's disorder.

Why do we care whether a treatment is specific? Is it not enough to know that it works? From a practical standpoint, we care because interventions are often costly and require extensive training and supervision of the interventionists. If patients benefit as much from regular meetings with a caring counselor as from a more elaborate and expensive intervention, then we have no need for a more complex intervention requiring extensive training. From a theoretical standpoint, we care because we wish to understand why change occurs, and we base our interventions on hypotheses about the critical changes processes. If our treatment fares no better than a basic control treatment, we need to question our beliefs about the necessary ingredients for change. Thus, tests of specificity speak to the issue of construct validity.

Finally, the investigator might wish to know which of two or more rival interventions is superior in the treatment or prevention of a given disorder. This is yet another form of the specificity question, and the most ambitious type of trial to mount because it requires large numbers of participants (see discussion on power in the section *Appropriate Statistical Analysis*). Such trials might involve the comparison of different schools of psychotherapy or of psychotherapy versus pharmacotherapy (Hollon, 1996; Mohr et al., 2009).

Special concerns arise when the comparison condition the investigator selects is treatment as usual (TAU) (Mohr et al., 2009; Nathan, Stuart, & Dolan, 2000). Such designs are especially common in effectiveness studies, in which investigators transport research clinic-tested treatments to clinical settings in the community. In such a design, the investigator does not dictate the contents of the comparison condition but relies on ordinary clinical care at the agency where the research is conducted or refers the participants randomized to TAU to community resources. Such studies have substantial external validity, in that they tell us whether the treatment in question is superior to what patients would otherwise have gotten, and they avoid the ethical problem of withholding treatment from someone who seeks it. However, internal validity problems can make their results hard to interpret: In many cases, the TAU group receives very little intervention, and any superiority of the new treatment may be due simply to the fact that the participants in that condition actually got treatment. In addition, the therapists in the new treatment may receive special training and supervision, creating the sort of excitement that can lead to a Hawthorne effect (Shadish et al., 2002). In other cases, there are problems in generalization of the results because the investigator fails to determine what sort of treatment the participants in the TAU condition received. Lacking this knowledge, how can we guess whether because the new treatment was superior to TAU in Clinic A it would be similarly superior in Clinic B where services are better than those of Clinic A? Only if the investigator has the resources to study TAU in a large number of clinics that represent adequately the range of services in the community, does it become possible

to conclude that on average the new treatment is better than typical practice.

Design issues for the investigator do not end with selection of the proper control group for the research question. To what degree should the investigator favor realistic conditions in the research versus tight control of factors such as amount of time in therapy? For example, patients in the community receiving psychotherapy for their depression generally meet with their therapists for 45 to 50 min at least once weekly. If they are in pharmacotherapy, however, after initial sessions they would rarely see their physician for more than 15-min sessions. Should the investigator constrain both treatment interventions to 45- to 50-min sessions even though this would not mimic real-world conditions and would be expensive in terms of the increase in physicians' time in the study? We have seen a change in this choice over our decades in the intervention field, where earlier researchers and reviewers favored carefully equating treatments on amount of time but, more recently, delivery of the treatments permits differences in amount of time and attention consistent with the way treatments are delivered in the field. As such, the field has moved from a more rigid adherence to maintenance of internal validity (by equating all conditions tightly on time in treatment) to permitting more focus on external validity (matching more closely what happens in the clinic) (Nathan et al., 2000). There is no right answer to this dilemma. Rather, the investigator must carefully consider the match of the design with the question he or she is attempting to answer and clearly describe any limitations in the interpretation of findings that follow from this choice in the discussion section of any report of the trial.

When two or more psychotherapy conditions are compared, the investigator must decide whether therapists are to be crossed or nested within treatments. When therapists are crossed with treatment condition, each therapist delivers all treatment conditions. When therapists are nested within treatment, each therapist implements only one of the treatment conditions. In the crossed case, the investigator can be sure that, if two treatments are found to differ, it is not because particularly good therapists were more likely to be assigned to one

treatment condition than another. However, the more the treatments are theoretically distinctive and require extensive training and commitment, the less likely it is that any one therapist can carry out two or more treatments with equal competence and commitment. For example, the committed cognitive therapist might not do a credible job of psychodynamic psychotherapy and might communicate to patients in psychodynamic therapy that they are receiving the less preferred treatment. The nested condition avoids this problem if the investigator recruits therapists who are equally skilled in and committed to their treatment approach for each condition—not an easy thing to do given the difficulty in assessing therapist competence reliably and validly (Waltz, Addis, Koerner, & Jacobson, 1993) and in finding skilled adherents of each treatment condition in many geographic areas. However, it leaves open the concern that the investigators may have unintentionally selected skilled therapists for their favored condition and less adequate ones for the other treatments.

ATTRITION

Attrition in intervention research is a common source of threats to both the internal and external validity of the findings. External validity is affected by a form of attrition that arises before potential participants ever begin the program. Who met criteria for entry to the study but chose not to participate? The larger the refusal rate, the less able we are to generalize the results of the study to the population of interest. The investigator can help the reader determine external validity in this sense by reporting how many suitable people were offered admission to the trial but refused and what reasons they gave for their rejection, perhaps using a flow chart such as that in Figure 28.1 to make attrition across the course of the trial easy to assess. Such a chart provides valuable information but does not reveal one important form of attrition: It is virtually impossible to determine how many potential participants learned of the procedures of the trial and decided against application for what might be important reasons that speak to a treatment's acceptability.

Once participants enter the study, the investigator faces the problem of maintaining their cooperation with the data-collection effort and retaining them in treatment. The longer the study, the more likely it is that participants will fail to continue with treatment, assessment, or both. Dropouts are a problem even in pretest–post-test designs, but attrition is especially problematic at follow-up when participants no longer have the incentive of receipt of the intervention to maintain their compliance with assessment procedures. The more people drop out of a research trial, the less representative the data are of the average person who enters treatment—a threat to external validity (Shadish et al., 2002).

Problems with attrition are even graver when attrition is differential (Shadish et al., 2002). Here internal validity is threatened. Differential attrition is obvious when dropout rates are higher in one intervention condition than another. It is more subtle, but equally dangerous, when different kinds of people drop out of one condition than another. Imagine two treatment conditions for which one condition is arduous but effective for those who stay the course, whereas the other condition is less demanding but also less effective. It is possible that for those people who complete treatment, the arduous treatment is more effective than the comparison condition, but that the less motivated people drop out of the demanding treatment. A comparison of treatment completers then yields findings that do not represent the results for the group of patients who started treatment: The investigators have lost the benefits of random assignment because there are now systematic differences between people in different treatment conditions (more motivated participants are in one condition than another), and internal validity is threatened. We will address the problems of attrition further when we discuss statistical conclusion validity. For now, will note only that the reader of any report based on completer rather than intention-to-treat analyses (Hollis & Campbell, 1999) should be wary and that investigators should make every attempt to continue to collect data on people who drop out of treatment (Lavori, 1992), for example, by using financial incentives to keep participants involved.

MODERATION AND MEDIATION

Researchers are usually not content to know simply whether or not a treatment works. They also want to know whether it works especially well or especially poorly for one type of participant or another and how a treatment works when it works. The first question concerns moderation and the second mediation (Baron & Kenny, 1986).

Moderation

Many psychotherapists believe it is important to select a treatment that matches a client in some important regard, and the search for client characteristics that predict better outcomes for one treatment than for another has become something of a quest for the Holy Grail in psychotherapy research. This sort of moderation effect, represented in statistical tests as a disordinal interaction, is not often found and is replicated even less often, perhaps because sample sizes are generally inadequate to detect such effects reliably or because they are truly rare. The failure of Project MATCH (Project MATCH Research Group, 1997) represents a sad example: This very large, very expensive, and well-conducted study was explicitly designed to test patient-treatment matching hypotheses in a sample of more than 1,700 patients with alcoholism. On the whole, previous findings from the moderation literature with alcohol-abusing or -dependent patients failed to replicate.

A different sort of moderation is the ordinal interaction. In one sort of ordinal interaction, the investigator may identify patient characteristics that are associated with poor treatment outcome relative to a control condition. For example, Fournier et al. (2010) found that antidepressant medication was significantly better than placebo but only for patients with severe depression. In such a case, we can say that treatment is efficacious but only for certain kinds of clients. Researchers are often less excited about this sort of interaction, in that it does not tell us what treatment would be better, for example, for patients who had less severe depression.

When moderator variables are selected after the trial has been conducted (at which point the researcher has knowledge of the results) rather than a priori, the investigator needs to be especially cautious in interpreting the findings, which may capitalize on chance (Kraemer, Wilson, Fairburn, & Agras, 2002). Until replicated, the results should be seen as heuristic for future research rather than definitive. For example, in the Treatment of Depression Collaborative Research Program (TDCRP), Elkin et al. (1989) conducted post hoc moderation analyses and reported that for patients with severe depression, cognitive therapy was not as effective as antidepressant medication. This effect was widely cited, and Elkin et al.'s caution that the moderation results were exploratory was often forgotten. Yet when DeRubeis, Gelfand, Tang, and Simons (1999) conducted a mega-analysis or patient-level meta-analysis¹ of four trials comparing cognitive therapy and antidepressant medication for severe depression (including the TDCRP), they found no evidence of superiority for medication over cognitive therapy.

According to some experts (Kraemer et al., 2002), the investigator should consider as potential moderators only those variables that can be assessed before treatment. In our view, this approach is too restrictive and risks overlooking important findings. For example, Addis et al. (2006) randomly assigned therapists in a health maintenance organization to training in cognitive-behavior therapy for panic disorder or to a waiting list for training. Panic disorder patients who saw the specially trained therapists improved significantly more than patients who saw the waiting-list therapists, but only when patients received at least eight sessions of treatment. When patients received fewer than eight sessions, there were no differences between conditions—a moderation effect. Note that in this study Addis et al. did not experimentally assign patients to fewer or more than eight sessions. Accordingly, this part of the report is quasi-experimental in nature and does not have the force of an investigation in which patients were randomly assigned to briefer or longer treatment.

¹In ordinary meta-analysis (Lipsey & Wilson, 2001), the researcher works with summary-level data (e.g., *Ms*, *SDs*, or effect sizes), whereas in mega-analysis or patient-level meta-analysis, the researcher works with the raw data from the original studies. As noted by Olkin (1995), this allows for more sophisticated analyses of the data in which the mega-analyst can examine hypotheses for which the original experiments did not provide analyses, such as the effects of individual patient characteristics and subgroup analyses.

Mediation

Analyses of mediation go deeper and test not whether patients have changed on clinical endpoints but what processes might account for that change.

In essence, tests of mediation are tests of causal agency or proposed mechanisms. What components of the larger treatment package played a causal role in bringing about the change in outcomes observed, and what mechanisms in the patient played a causal role in transmitting the impact of treatment to the outcomes of interest? Such tests speak to the construct validity of the research. For example, it goes without saying that cognitive therapists believe that changing patients' maladaptive cognitions is the central curative process in psychotherapy. In treatment research, mediation is tested by determining whether changes in the outcome variable over time or differences in treatment efficacy between two treatment conditions (e.g., cognitive-behavior therapy versus a placebo) can be statistically accounted for by changes in the mediator (e.g., cognitive change) (Kraemer et al., 2002). For example, Smits, Rosenfield, McDonald, and Telch (2006) collected data across 15 exposure trials from patients in treatment for social phobia, assessing cognitive variables before each exposure trial and fear after each trial. The authors demonstrated that reductions in patients' predictions that they would look anxious, stupid, and so forth statistically accounted for reductions in fear across trials. A critical factor in the appropriate test of mediation is the researcher's ability to demonstrate that change in the mediator preceded change on the outcome variable. Otherwise, the mediator could just as easily have been caused by the outcome as the converse or be a time-varying third variable not causally related to the outcome at all. We will return to this point with a specific example in the Alliance section.

STATISTICAL CONCLUSION VALIDITY

Type I and II Errors

Perhaps because the opportunity to collect data in an intervention trial is precious, investigators tend to want to collect data on many variables, including outcome variables. This is understandable but runs the risk of Type I error. Investigators have been

known to test 20 or more outcome variables, all with an alpha level of .05, and then interpret a smattering of significant effects as indicating they have demonstrated efficacy for the treatment tested.

Worse yet, the investigator might report only those measures that yielded findings consistent with the hypothesis. After the fact, it is tempting to conclude that the measures yielding significant outcomes are the most valid measures.

Using a large number of outcomes and then adopting stringent Bonferroni-corrected alpha values is not a reasonable strategy for intervention research, in which each additional participant is costly and difficult to recruit (Nakagawa, 2004). The necessary sample sizes for testing large numbers of outcomes with experiment-wise corrections to *p* values are unlikely to be achieved. Rather, the investigator should select the best measure(s) as the primary measure(s) of outcome (a judicious number of other variables can still be tested as secondary outcomes) or seek to cut the number of outcome variables by reducing the data before hypothesis testing (Rosenthal & Rosnow, 2008), for example, by conducting a principal components analysis of the outcome variables and using the component scores rather than individual variables for hypothesis testing. Kraemer et al. (2003) have provided suggestions for combining data from multiple informants, the sort of data common in research with children.

The problem of Type II error also plagues intervention research. If a treatment is genuinely effective, most often it will be found to be statistically significantly different from a waiting-list condition. Comparisons with an active treatment are another matter, however, because effects sizes in this case are likely to be no more than medium. Psychotherapy research trials are typically underpowered to detect this sort of difference (Kazdin & Bass, 1989), and researchers and readers frequently make the mistake of concluding there is no difference between treatments when from the outset the sample size precluded differences of a medium size from being statistically significant. With increased awareness of this problem, researchers have banded together across centers to conduct multisite trials to obtain the sample sizes they need to test important

questions of differential efficacy. This, of course, introduces another set of challenges in ensuring that methods are comparable across two or more sites (Kraemer, 2000).

Noninferiority and Equivalence Testing

Often in intervention research, the investigator's goal is not to test whether one intervention is better than another but rather to test whether some new intervention is as good as a well-established treatment. Perhaps this new intervention is briefer, less costly, or easier to implement than the standard treatment. A frequent mistake is for the researcher to conduct standard tests for statistically significant differences between interventions' impact, find none, and declare that the two procedures are comparable in their efficacy. This is not correct, however; they were just not statistically significantly different in this trial. To conduct a proper test of the investigator's hypothesis requires a test of noninferiority (the new intervention is no worse than the standard intervention) or of equivalence (the new intervention is neither worse than nor better than the standard intervention) (Stegner, Bostrom, & Greenfield, 1996).

To conduct such tests, the investigators must identify in advance what effect they consider so small as to be clinically irrelevant (Piaggio, Elbourne, Altman, Pocock, & Evans, 2006). It is best if the field has some consensus on how small that difference should be and on what outcome variable it should be tested. The investigators then conduct statistical tests for equivalence or noninferiority (Stegner et al., 1996), depending on which fits best with their research question. These tests involve rejecting the null hypothesis that the two interventions are different in their efficacy rather than the usual null hypothesis of no difference. Equivalence and noninferiority trials have become common in biomedical research but not in psychological research, in part because the large sample sizes required for this approach are unlikely in psychotherapy research. Some large-scale intervention or prevention trials might have the requisite sample sizes. Because sloppy research can lead investigators to find no difference between treatments when such a difference does exist (error swamps the treatment

effect), equivalence and noninferiority trials must be exquisitely conducted with an eye to internal validity for the reader to have confidence in their conclusions.

Effect Size and Clinically Significant Change

A common error for researchers and research consumers is to state that the treatment effect was really big because the p value was very small. Despite decades of exhortations that psychologists should report and interpret the effect sizes derived from their findings (e.g., Wilkinson & Task Force on Statistical Inference, 1999), researchers and consumers of research still confuse the interpretation of a p value with the size of the effect. Because the p value is highly dependent on not only the size of the effect but also the sample size, we cannot determine how big the difference was between two intervention conditions by a report of the p value alone. To give the research consumer an idea of how big the difference is in a way that is not dependent on the scale of the particular outcome measure used, the researcher should report an effect size, which is commonly given in standard deviation units, although this varies with the type of statistical analysis (Cohen, 1988; Kraemer & Kupfer, 2006).

Even the effect size, however, can be hard to interpret, as changes will be bigger on measures that are more reliable (because the standard deviation will be smaller). Patients might change substantially according to average effect sizes on a very reliable measure but still be quite impaired. Accordingly, if the investigator wishes to report how many of the patients are doing well, it is best to incorporate an additional approach. These approaches might include reporting what percentage of patients in each group no longer meet diagnostic criteria for the primary diagnosis at the end of treatment or reporting what percentage of patients in each group meet criteria for clinically significant change (Jacobson & Truax, 1991). The calculation of clinically significant change, as formulated by Jacobson and Truax, includes determining whether a given patient has changed to a statistically reliable degree, and whether she or he is more likely to fall in the distribution of the scores of a normal sample rather than

a sample of people with the disorder under investigation.

Research groups in some areas have developed conventions for what they consider recovery in their field that are based on yet different criteria. For example, researchers in major depression have adopted a set of criteria for the definition of recovery from a major depressive episode (Frank et al., 1991). Whatever method researchers choose, they may then test whether the intervention groups differ on the percentage of patients recovered, although this test will generally be less powerful than tests of the continuous variables.

Once some metric of success has been established, the investigator may then report the results using an effect size from biomedical research that has become increasingly popular in psychiatric research, that is, *number needed to treat* (NNT). NNT is the number of patients one would have to treat with a given approach to achieve one more successful case than if an equal number of patients received the control or alternative intervention. The smaller the NNT, the larger the treatment effect size. For example, the results of meta-analysis of psychotherapy versus waiting-list control groups recast as NNT by Kraemer and Kupfer (2006) suggested that for every 3.1 patients who received psychotherapy rather than being on a waiting list, one more patient would improve. Kraemer and Kupfer (2006) described methods for converting continuous measures of outcome into NNT as well.

Appropriate Statistical Analysis

Standards for the statistical analysis of data from intervention research are ever higher, and readers of older research papers may find procedures that would lead a manuscript to be rejected for publication in the 21st century. Several common issues arise.

First, as noted, missing data are common in longitudinal research. Patients drop out of treatment, and participants fail to complete all assessment points. The investigator who conducts only *completer analyses* of people who provide all data and who complete treatment is analyzing data that are unrepresentative of the sample that was randomized to intervention conditions. Once awareness of this

problem developed, researchers began to use a procedure called *last observation carried forward* (LOCF) to cope with missing data (Shao & Zhong, 2003). For example, if the client drops out of treatment, refuses post-test assessment, or both, her or his pretest score (or most recent score available) is substituted for the missing datum at post-test.

Although this may be a better approach than a completer analysis, it may underestimate treatment effects, in that people sometimes disappear from trials because they are doing well and feel no further need for assistance. (See Lavori, 1992, for a further critique of the LOCF approach.) Accordingly, the development of more sophisticated approaches to missing data has been a boon to intervention research. The approach variously called hierarchical linear modeling, random regression modeling, and multilevel modeling allows the investigator to model the trajectory across time on the outcome variable for each subject on the basis of however many data points that subject completed (Bryk & Raudenbush, 1987). The trajectories for the individual subjects then become the data for between-group analysis. If data are missing at random or missing completely at random, multilevel modeling approaches can cope well with less than complete data (Hedeker & Gibbons, 1997). To get the greatest benefit from this approach, the investigator would do well to collect outcome data at multiple points across the trial rather than just at pretest, post-test, and follow-up. Multiple data points not only provide greater statistical power (de Jong, Moerbeek, & van der Leeden, 2010) but also allow a better determination of the individual trajectories and is especially helpful in the case of dropout.

Multilevel modeling allows the investigator to address readily a second common statistical problem in intervention research, that of nesting. Usually a given therapist sees multiple patients in a trial, and to the degree that there are any differences across therapists in how effective they are, patients seen by a particular therapist share some variance in their outcomes that is due to the therapist. Other forms of nesting include the group in trials of group therapy, the classroom and the school in interventions conducted in education settings, and the site in multisite trials. When nesting is ignored in the analyses, Type

I error can increase dramatically (de Jong et al., 2010). Multilevel modeling allows the researcher to build such nesting into the statistical analysis. A particularly thorny issue in studies involving nesting is whether to treat therapists, for example, as fixed or random effects. To treat therapists as fixed effects in the analyses means that the investigator cannot generalize the results of the trial beyond these specific therapists—a point largely ignored in the treatment literature. If therapists are treated as a random effect, it is reasonable to generalize the results of the study to similar therapists with comparable training (Cris-Christoph, Tu, & Gallop, 2003). Because inclusion of therapists as a random effect diminishes statistical power, investigators are loath to do this (Cris-Christoph et al., 2003).

The third common problem we will mention arises when full randomization is not possible because the investigator is conducting quasi-experimental research or when the investigator randomizes intact groups (e.g., schools) to an intervention program. In such cases, it is likely that there will be pretreatment differences between intervention conditions. Given large enough sample sizes, the investigator can overcome the problems represented by the preintervention differences with propensity score analysis (Joffe & Rosenbaum, 1999). In this approach, logistic regression is used to predict treatment group assignment by all possible covariates to yield a predicted probability of being in one group versus the other. This resulting probability score can then be used as a single covariate in the analyses of outcome. Alternatively, the propensity score may be used to form a number of strata within which participants are fairly closely matched. In this case, the analyses are conducted within strata. The latter approach clearly requires a large sample to execute with adequate power. If the investigator has collected data on all the important covariates, then propensity score analysis is effective in diminishing selection effects (Shadish et al., 2002).

CONSTRUCT VALIDITY OF AN EXPERIMENT

Construct validity refers to the extent to which one can make inferences from the procedures of a study

to the higher order constructs they were intended to represent (Shadish et al., 2002). With respect to randomized controlled trials, that means the extent to which the investigator has accurately identified the causally active components of the treatment manipulation. Most treatments consist of multiple components, some specified by theory and others common to the treatment enterprise. Construct validity asks whether we understand the causally active components of the intervention and the mechanisms through which they operate.

As previously described in the section on selection of control groups, different types of control groups allow for increasingly greater confidence that the theory as specified was adequately tested. For example, cognitive therapy is predicated on a theory that states that teaching patients to examine the accuracy of their beliefs should reduce distress. If we were to compare cognitive therapy to its absence we might very well produce an internally valid difference that could be attributed to the experimental manipulation (and thereby establish its efficacy), but it would not do much to convince a skeptic that it was the targeting of inaccurate beliefs that led to the subsequent reduction in distress. It could just as well have been the nonspecific aspects of the treatment package (expectation for change and personal contact with the therapist) that were responsible for the change observed. Construct validity would be enhanced if cognitive therapy were found to be superior to a nonspecific control that was equated for the mobilization of expectations and therapist contact. Construct validity would be enhanced even further if change in cognition was linked to subsequent change in distress in a manner suggestive of causal mediation. The issue then with respect to construct validity is whether the experiment conducted provides an adequate test of the underlying substantive theory. To do so it must implement the treatment in question in an adequate fashion and control for other alternative explanations not specified by theory.

Treatment Manuals

Treatment manuals represent an attempt to specify the underlying principles and specific behaviors that together constitute a treatment intervention.

Manuals differ in the extent to which they constrain the therapist's behavior: Some provide considerable latitude for the clinician to respond to the specific needs of the client as they unfold over time, whereas others go so far as to specify the actual dialogue that the therapist is supposed to follow across the course of treatment (Luborsky & DeRubeis, 1984). Treatment manuals serve a useful purpose in communicating what is done in a given intervention and facilitate dissemination to other researchers and therapists in other settings. They also reduce variability between therapists within a trial (Crits-Christoph et al., 1991), thereby enhancing the extent to which the essential aspects of a treatment are implemented (construct validity) and increasing the likelihood that treatment differences will be detected relative to controls (statistical conclusion validity).

Nonetheless, treatment manuals are neither necessary nor sufficient to ensure that an intervention has been adequately implemented in a given trial. In the classic Temple study, experienced psychoanalytic therapists operated without a formal treatment manual but were able to instantiate dynamic treatment in a representative fashion nonetheless, due no doubt to years of training and their experience in analysis (Sloane, Staples, Cristol, Yorkson, & Whipple, 1975). At the same time, despite training to a manual, the National Institute of Mental Health (NIMH) Treatment of Depression Collaborative Research Project found that therapists with prior experience with cognitive therapy got considerably better results than those who did not have prior experience (Jacobson & Hollon, 1996). To assure the research consumer that the treatment was faithfully and well conducted, additional steps are required of the investigator. We turn to those next.

Therapist Adherence

Adherence refers to the extent that the therapists implement the therapy as intended. A study would have little construct validity if therapist behavior bore no relationship to what was intended by theory. Adherence can be measured on the basis of therapist self-report or the completion of postsession checklists, but the preferred manner is by actual direct observation of the session itself, often in the form of ratings of audio or video tapes

(Chevron & Rounsaville, 1983). Such methods require observers trained to recognize the behaviors specified by theory but do not necessarily require that those observers be competent to implement the therapy themselves. For example, Hill, O'Grady, and Elkin (1992) found that cognitive therapy could readily be distinguished from interpersonal psychotherapy and each from the clinical management component of pharmacotherapy by nonprofessional raters listening to audiotapes from the TDCRP.

Therapist Training and Competence

Competence refers to the extent to which the therapists perform the intended therapy in a skillful fashion. Competence is related to adherence (you cannot perform a therapy well if you are not performing the therapy) but can be differentiated at least in theory: It is possible to implement the various components of a treatment in a highly recognizable manner that is neither skillful nor responsive to the needs of the patient at a given moment. To use an analogy, it is possible to play a musical piece in a manner that others could recognize without doing so in a manner that is pleasing to the ear. In the TDCRP, adherence unexpectedly functioned as a suppressor variable, the inclusion of which enhanced the relation between competence and outcome in cognitive therapy, suggesting that at least some of the therapists were adherent to the approach without being all that competent in its execution (Shaw et al., 1999). Like adherence, competence can be rated in a variety of ways but most often is rated by expert therapists working from tapes of actual sessions.

Therapist training is intended to enhance both adherence and especially competence. Just how much training is required and how best it is accomplished is a matter of some debate, but it seems fair to say that the strategies pursued should match the purposes of the study. In the typical efficacy trial, the goal is usually to determine whether a given treatment has a causal impact under ideal conditions. In such a trial, it seems reasonable to ask that the therapists be trained to the point at which they can implement the therapy in a manner specified by theory. In other types of trials, especially some types of effectiveness studies, the goal may be to see how

much change can be produced by therapists trained to whatever level of competence is allowed given the pragmatic constraints in the natural environment. Either level of training is fine so long as it is clear what was going on, and causal inferences are drawn in an appropriate manner.

Training is often supplemented by subsequent supervision that may continue through the duration of the trial. Again, just how to supervise and how much supervision to provide should be determined by the questions being asked in the particular trial, and problems arise only when the inferences drawn do not match the implementation. For example, in the TDCRP, therapists with varying levels of prior experience with and commitment to cognitive therapy were provided with several days of training and intensive supervision while working with several practice cases each. Once the study proper started, supervision was cut back unless ratings on a competence measure dropped below a preset cutoff (Shaw, 1984). What the investigators found was that rated levels of competence dropped from the training phase into the study proper (Shaw et al., 1999) and that cognitive therapy was less efficacious than medications and no better than pill-placebo for patients with more severe depressions (Elkin et al., 1995). By way of contrast, in a subsequent trial DeRubeis et al. (2005) selected experienced cognitive therapists at one site and, at a second site, continued intensive training throughout the course of the study proper for the less experienced cognitive therapists at that site. The investigators found that cognitive therapy was as efficacious as medications and superior to pill-placebo in the treatment of patients with depressions of comparable severity. Although both trials were intended to speak to the efficacy of cognitive therapy as specified by theory, they differed considerably in the nature of the therapists who they selected and the quality and intensity of the training and supervision that they provided—differences that appear to be reflected in the outcomes that they generated.

Allegiance

Investigator allegiance is an important correlate of variance in outcomes across the treatment literature (Luborsky et al., 1999). Treatments usually do

better in the hands of investigators who are invested in their outcomes, and comparisons between different treatments often rise and fall on the basis of who carries out the study. There are at least two possible interpretations. First, it may be that investigators with a vested interest cannot or will not conduct a fair trial and that bias (untended or otherwise) colors the results. Second, it could be that some investigators are simply more competent when it comes to implementing some treatments than others and that differential outcome across studies reflects differential competence of the investigators. For example, cognitive therapy was found to be superior to antidepressant medications in a study conducted at the site where the psychosocial treatment was first developed (Rush, Beck, Kovacs, & Hollon, 1977). However, pharmacotherapy was not adequately implemented: Dosages were marginal, and tapering was begun before the end of treatment. Subsequent studies that did a better job of implementation typically found cognitive therapy as efficacious as but not better than medication (e.g., Hollon et al., 1992).

Unintended bias can be addressed by adherence to principles of good research design (random assignment and blinded evaluators), but differential competence can be resolved only by including investigators in the research team who have expertise (and are invested) in each of the modalities tested. This is the principle of adversarial collaboration that has been described in the cognitive psychology literature to offset the operation of bias (Mellers, Hertwig, & Kahneman, 2001). It is important that this allegiance and expertise permeate all aspects of the study, starting with the investigator team and including the therapists who provide the actual interventions.

Expectancy

Patients enter treatment in the hope that it will make things better, and expectations of improvement have been shown to have a powerful effect on outcomes in their own right. Some treatments do a better job of mobilizing expectations for change than others, and differences in expectations can influence the comparisons between conditions. To the extent that the mobilization of expectations can

be considered an inherent part of a given intervention, then expectancies pose no threat to the internal validity of the design: An effect is an effect and can be attributed to the intervention regardless of how it was produced.

Isolating the contribution of expectancy effects, however, can play a major role in determining whether a treatment works for the reasons specified by theory, a matter of construct validity. For example, it is routine to test new medications against a pharmacologically inert pill-placebo that controls for all the psychological aspects of medication-taking, including the expectation of change. Only if the novel medication exceeds that pill-placebo control in a beneficial fashion is it allowed to be marketed to patients. The pill-placebo control is presumed to generate similar expectations for relief to those generated by the actual medication (as well as other nonspecific benefits of contact with a treating professional), and any differences observed are presumed to be the consequence of the pharmacological properties of the medication. Similar steps are often taken in psychotherapy research to determine whether comparison conditions are equated for the expectations that they generate at the outset of treatment, and studies are sometimes criticized for including *intent-to-fail* controls if they cannot be shown to be equated for initial expectations (Baskin, Tierney, Minami, & Wampold, 2003). We suspect that expectancy differences may be one result of differences in allegiance to different treatments. That is, if investigators compare some alternative treatment to their preferred treatment, they may inadvertently do so in a way that communicates to patients and therapists which treatment is expected to be inferior (e.g., by their enthusiasm, the quality of the training materials, and supervision provided).

Alliance

Patients not only have expectations regarding treatment outcomes but also form relationships with their therapists. The quality of the working alliance represents one attempt to operationalize the quality of the therapeutic relationship, and the term is often used in a generic fashion to refer to the larger construct of relatedness between patient and therapist (Goldfried, 1980). It has long been noted that

various measures of alliance predict treatment outcome (Horvath & Symonds, 1991). What is not so clear, however, is whether the therapeutic alliance plays a causal role in producing those outcomes. In most instances, alliance is rated periodically across treatment and correlated with treatment outcome. This means that early change in symptoms could be driving the quality of the relationship rather than the other way around.

In a pair of studies, DeRubeis and colleagues tested this hypothesis by assessing symptom change both before and after therapy sessions that were rated both for techniques specific to cognitive therapy and for nonspecific quality of the therapeutic alliance (DeRubeis & Feeley, 1990; Feeley, DeRubeis, & Gelfand, 1999). They found that after they controlled for prior symptom change, adherence to cognitive therapy in early sessions predicted subsequent change in depression, whereas ratings of the therapeutic alliance did not. Moreover, rated quality of the therapeutic alliance improved over time as a consequence of prior symptom change. These findings suggest that for cognitive therapy (as practiced in their samples), adherence to the techniques specified by theory drove subsequent symptom change and that change in turn drove the rated quality of the therapeutic alliance. At the same time, other studies that have controlled for prior symptom change in a similar fashion have found that rated quality of the alliance does sometimes predict subsequent symptom change (Klein et al., 2003), although effect sizes are lower than in studies in which the temporal sequence problem was ignored. What seems likely is that the therapeutic alliance may be either a cause or a consequence of improvement in treatment and that it is important to control for the temporal relations between treatment process and symptom change whenever investigating such effects.

Exclusion of Medication or Medication Stability

One issue that is sometimes confronted in psychotherapy research is what to do about patients who enter a study already on medications. On the one hand, patients often are reluctant to give up medications that they are already on, and excluding those

patients from the trial would reduce the external validity of the design. On the other hand, patients are taking medications precisely because they think they make a difference, and to the extent this is true, such effects can obscure (or facilitate) the effects of psychotherapy. This problem is especially common in the treatment of anxiety and panic disorder in which case patients often are quite unwilling to make any changes in long-standing medication patterns.

Some investigators resolve this dilemma by asking patients to discontinue any psychoactive medications before entering the trial (and losing at least some potential patients who refuse to do so), whereas others are willing to allow patients to remain on stable doses of existing medications. Either strategy represents a reasonable accommodation to the practical realities of conducting clinical trials with patients who can choose to not participate, and causal inferences can still be drawn so long as they are tempered by recognition of the possible influences of concurrent medication usage. In such situations, it is particularly important to monitor what medications patients are taking and to conduct secondary analyses that control for medication usage. Changes in medication usage that occur after randomization are particularly problematic because they could be a consequence of differential treatment. Early comparative trials often found that the beneficial effects of adding psychotherapy were obscured by the fact that medication doses typically were lower in combination treatment (Hollon & Beck, 1978).

Exclusion of Other Treatment

Similar issues are raised by allowing patients to pursue other psychosocial treatment during the course of the study proper. As was the case for off-protocol medication treatment, excluding patients who refuse to give up ongoing psychotherapy threatens the external validity of the design, whereas allowing such patients in the trial presents problems for construct validity (if comparable in nature across conditions) or internal validity (if differential across the conditions). Moreover, having patients in two different kinds of psychotherapies risks having them working at cross-purposes. What many investigators

do is to ask that patients discontinue any outside psychotherapy directed at the disorder under study for the duration of the trial, but they often make exceptions for psychosocial interventions directed at other issues, such as marital or family therapy.

Adequacy of Dose of Treatment

The essence of construct validity is that the underlying constructs are tested in a manner specified by theory. That suggests that what constitutes an adequate dose of treatment is determined by the question being addressed. There is nothing inherent in any given dose of treatment, but the doses selected for testing ought to be consistent with what is specified by theory. For example, neither the early study that compared cognitive therapy to inadequate doses of medication treatment (Rush et al., 1977) nor the subsequent NIMH TDCRP that left the implementation of cognitive therapy in the hands of relatively inexperienced therapists (Elkin et al., 1989) provided an adequate basis for comparing the relative efficacy of the two modalities as each is ideally practiced. Meta-analytic reviews suggested that each is comparable to the other when optimally implemented (Cuijpers, van Stratten, van Oppen, & Andersson, 2008). However, those same meta-analytic reviews also showed an advantage for full-strength medication treatment over relatively abbreviated courses of psychotherapy in managed care settings. It seems fair to conclude from these reviews that medication treatment is superior to psychotherapy when the latter must be restricted in dosage because of pragmatic constraints so long as one does not conclude that that relative inferiority reflects anything other than those pragmatic constraints.

EXTERNAL VALIDITY

External validity refers to the extent that treatment outcomes observed in controlled trials can be generalized to populations and settings of interest (Shadish et al., 2002). Several aspects of generalizability need to be considered, including but not limited to variation across patients, therapists, and settings. External validity is closely related to the notion of clinical utility and sometimes is

considered in conjunction with cost (American Psychological Association [APA], 2002).

Continuum of Efficacy-Effectiveness Research

It has become commonplace in recent years to differentiate between efficacy and effectiveness research (Nathan et al., 2000). According to this distinction, efficacy research is said to be conducted in highly controlled settings using carefully selected patients who are randomly assigned to treatment by highly trained therapists. Conversely, effectiveness research is thought to be conducted in real-world settings in which random assignment to differential treatment is not always feasible and in which presumably more complicated patients are treated by less experienced therapists working under pragmatic constraints imposed by their many clinical demands (Seligman, 1995). Although there is some truth to these perceptions, we think it is unwise to draw too sharp a distinction between efficacy and effectiveness research and prefer instead to think in terms of internal and external validity (Chambless & Hollon, 1998). Any given study can be evaluated with respect to how it scores on each dimension. The goal is to determine how well a given intervention works in the real-world settings in which it needs to be applied to wholly representative patients.

Treatment outcomes are largely a function of two sets of factors: patients and procedures. Patients' characteristics are fixed at the time they first present for treatment, but procedures (including therapist skills and setting considerations) can be changed if there is sufficient reason. That is why we think that it is especially important to conduct research on wholly representative patient samples. If a treatment does not work for the patients for whom it is intended, then there is little that can be done to improve the situation. There can be great value in first establishing that something works under ideal conditions even if it does not initially generalize well to real-world settings, however, because therapists can be better trained and contexts modified if there is sufficient reason to do so. That is why we emphasize internal validity over external validity and selecting representative patients (rather than therapists or settings) in early trials.

Generalizations

To clinical settings. It is important to know whether treatments that are efficacious in research settings generalize to the kinds of applied settings in which most patients are treated. Applied settings can vary with respect to the caseload expected, the length of treatment that is feasible, and the demands on clinician time (Weisz & Addis, 2006). For example, we were once informed by therapists who worked for a large managed care organization that they could see patients as often as they thought appropriate, just so long as they started six new patients a week. Although no constraint was placed on the number of sessions that could be offered to any given patient, the pressure to add so many new patients effectively limited the number of sessions that could be provided.

To patients in community settings. There is a widespread perception that patients treated in efficacy studies in research settings are necessarily less complicated or comorbid than patients found in community settings (Westen, Novonty, & Thompson-Brenner, 2004). Although this may have been the case in early analogue studies, there is no necessary reason why complex and comorbid patients must be excluded from controlled trials. For example, two thirds of a sample of patients selected on the basis of meeting criteria for major depressive disorder in a recent efficacy trial conducted in a research clinic met criteria for one or more additional Axis I disorders and half met criteria for one or more Axis II disorders (DeRubeis et al., 2005).

Conversely, when Stirman, DeRubeis, Crits-Christoph, and Brody (2003) matched information found in community outpatients' charts to the inclusion and exclusion criteria used in published trials, they found nearly 80% of the diagnosed patients in community settings would have qualified for inclusion in one or more RCTs. The major reason patients would have been excluded from RCTs was not that they were too complicated or comorbid. Rather, their conditions were not severe enough, in that many carried diagnoses of adjustment disorder. Such disorders are by definition transient and will likely never warrant the conduct of an RCT. It is likely that at least some of these patients would have

met criteria for Axis I or Axis II disorders if diagnosed according to strict research criteria; clinicians in applied settings often give diagnoses of adjustment disorder to avoid stigma. If so, then an even larger proportion of the patients seen in applied settings would likely qualify for inclusion in one or more controlled clinical trials. Clearly, research needs to be done with the kinds of patients found in applied community settings, but it is not clear that they will prove to be all that different from patients found in clinical research sites.

To clinicians in community settings. Clinicians in community settings are likely to be less experienced with specific disorders and less highly trained than clinicians in research settings. It is an open question just how well these clinicians can implement the kinds of treatments developed in research settings and tested in efficacy studies. On the one hand, there is little evidence that experience or professional status necessarily guarantees superior performance (Jacobson & Christensen, 1996), but on the other, much of the variability in outcomes across different studies appears to be related to the competence with which the therapists can perform the given interventions (Jacobson & Hollon, 1996). It is likely that years of experience and professional status are not particularly good markers of competence with a particular approach.

It is not clear just how much training is required to allow community clinicians to perform with the same level of proficiency as research therapists or how feasible such training is to provide in community settings. What does appear to be clear is that the provision of treatment manuals and brief workshops are not sufficient to help community therapists reach a reasonable level of proficiency (Miller, Sorensen, Selzer, & Brigham, 2006). The same concerns apply to pharmacotherapy as practiced in the community; surveys suggest that typical practice often falls far short of what gets done in controlled trials in research settings (Trivedi et al., 2004).

To patients of diverse backgrounds. It is important to include a diverse array of patients with respect to race, ethnicity, and socioeconomic status (SES) in research trials to determine whether study findings generalize to such patients. Minority

patients not only tend to be few in number in clinical trials (just because they are fewer in number) but also often are under-represented because they are suspicious of the motives of clinical researchers (with considerable historical justification). Language can be a barrier for many ethnic patients, and low SES patients often face barriers to participation with respect to transportation and child care. It is often the case that special efforts have to be made to recruit and retain such patients in controlled trials (Miranda et al., 2003).

To other research settings. Treatments typically are developed by a single individual or group and tested in the sites at which they were first developed. This means that early studies typically are done by groups that are invested in the treatment they are testing. Replication is a key principle in science, and if a treatment is truly efficacious, then other investigators at other sites should be able to replicate the findings generated in those initial trials. Some shrinkage is to be expected, as it is unlikely that other groups will be as expert in a given modality as the people who developed it. Nonetheless, if a treatment is to have value in real-world applications, then it must perform well in the hands of other groups at other sites. At the same time, the principle of replication implies that those other groups put in a good faith effort to learn how to implement the treatment in a reasonable fashion. Replication means that we should be able to reproduce the earlier findings if we implement the same procedures: With respect to treatment trials, this means that we must implement the treatment in question with reasonable fidelity.

Feasibility

Feasibility refers to the ease with which treatment can be delivered to actual patients in the settings in which they are typically treated (APA, 2002). Feasibility incorporates such factors as acceptability, patients' willingness and ability to comply with the requirements of the intervention, ease of dissemination, and cost-effectiveness. We examine each set of issues in turn.

Acceptability and patient-preference designs. For a treatment to be applied in actual practice, it must

be acceptable to potential patients and also to other relevant parties, including therapists and administrators. There are many reasons why individual patients may prefer not to receive particular treatments, and these preferences need to be respected (APA, Task Force on Evidence-Based Practice, 2006). At the same time, patients often have trouble accepting the very interventions that are most likely to be useful to them. For example, exposure therapy is the best established psychological treatment for obsessive-compulsive disorder, yet it is frightening to most patients because it calls for them to do exactly those things that they most fear. Similarly, patients often refuse to take monoamine oxidase inhibitors because of the dietary restrictions required to avoid the risk of hypertensive crisis, despite the fact that these medications often represent the best pharmacological option for patients who are refractory to easier-to-manage medications.

Therapists also may be reluctant to implement treatments that they find threatening or distasteful, and some go so far as to ignore empirically supported interventions in favor of more traditional approaches that they find more compelling or in which they are already trained. Administrators also may play a role. We recently encountered a situation in which the Office of Legal Affairs tried to prevent a therapist from conducting exposure therapy with a competitive swimmer who had developed a morbid fear of drowning because it wanted to limit the legal exposure of the university in the unlikely event that the patient drowned during treatment.

Acceptability also plays a role in the decision to randomize. Given that patients differ in terms of how they are likely to change over time in the absence of treatment, we have argued in the section on Randomization that random assignment to intervention conditions is essential if any differences observed in outcomes are to be attributed to the experimental manipulation. Something like hormone replacement therapy provides a cautionary tale regarding what happens when we do not randomize and allow patients characteristics to determine treatment: Prognosis gets misconstrued as a treatment effect (Chlebowski et al., 2003; Manson et al. 2003).

That being said, no one likes to be randomized, and no one likes to randomize someone else to treatment. Prospective patients dislike the notion of leaving their fates to chance, and clinicians typically assume, rightly or wrongly, that they have the experience and expertise to select the best treatment for a given patient. Some have gone so far as to say that randomization may turn out to be worse than useless: Choice is itself curative, and patients may adhere better if they get the treatment they prefer (Seligman, 1995). A number of potential solutions have been proposed, from relying on retrospective surveys of patient satisfaction to the use of quasi-experimental designs that attempt to control for some of the threats to internal validity, to the use of benchmarking in which the results of open trials in applied settings are compared with those obtained in randomized controlled trials. Retrospective surveys are wholly uncontrolled and highly susceptible to the risk of confounding patient characteristics with treatment effects, quasi-experimental designs provide some protection against internal validity threats but that protection is only partial, and benchmarking is interpretatively ambiguous because any similarities or differences observed could be the product of differences in patients or procedures.

These concerns have led to an interest in the patient-preference design. Brewin and Bradley (1989) proposed a comprehensive cohort design that restricts randomization to only those patients who are willing to accept it; those who refuse are provided with the treatment they prefer. Others have proposed a two-stage randomization design in which patients are randomized to be (a) randomized to condition or (b) allowed to pick their treatment (Wennberg, Barry, Fowler, & Mulley, 1993). Neither approach is wholly satisfying because causal inferences can be drawn with any confidence only for patients who were randomized to the respective treatment conditions. The comprehensive cohort design does allow randomized patients to be benchmarked against those who refused and got their preferred treatment, and the two-stage randomized design does allow causal inferences to be drawn about the effects of patient choice. It remains unclear just how great a problem we are dealing with. A systematic meta-analytic review of

patient-preference designs found that patient preferences led to substantial rates of refusal; prospective patients who are employed and well-educated are especially likely to refuse randomization (King et al., 2005). This suggests that patient preferences are a threat to the external validity of the typical RCT. At the same time, there was little evidence that patient preferences compromised the conclusions drawn from the studies: Differences in outcomes between randomized and preference groups typically were small (especially in larger studies) and when they were evident in smaller trials, were inconsistent in direction. The preference problem may not be as big as some believe.

The recent NIMH-funded Sequenced Treatment Alternatives to Relieve Depression (STAR*D) project used a particularly interesting strategy called equipoise stratified randomization in which patients or clinicians were allowed to rule out treatment strategies that they found unacceptable. Patients were subsequently randomized to the remaining options and only included in analyses that compared patients who accepted randomization to a given option (Lavori et al., 2001). For example, patients who showed a partial response to their initial medication could choose to not be switched, and patients who were unable to tolerate their initial medication could choose to not have their medication augmented. This approach appears to be the most compelling of the patient-preference strategies and has generated considerable enthusiasm in the field. Still, it is not without its problems. Rates of refusal were suspiciously high for some conditions (less than a third of the participants were willing to be randomized to cognitive therapy): Permitting patients to opt out of specific strategies might encourage greater rates of refusal than might otherwise have occurred. Moreover, treating clinicians often have even stronger preferences than their patients, which may inflate rates of refusal. Nonetheless, equipoise stratified randomization is an interesting approach that warrants further consideration.

Compliance by clients with treatment requirements. Even the most efficacious treatment will not work if the client does not implement the necessary steps. This issue is closely related to

acceptability but is slightly more subtle. Patients may readily accept a treatment but not necessarily carry through with all the steps required to derive maximum benefit. For example, compliance with homework tends to predict subsequent improvement in cognitive therapy for depression (Burns & Spangler, 2000). Similarly, it would be inappropriate to say that a medication does not work if a patient does not take it. From a methodological standpoint, it is important to assess compliance so that accurate conclusions can be drawn with respect to treatment efficacy. For example, in a recent follow-up design, we monitored medication compliance in patients assigned to continuation treatment and conducted secondary analyses in which we censored patients who were less than fully compliant in the several weeks before a relapse (Hollon et al., 2005). Whereas the primary analysis that left such relapses uncensored provided the best estimate of how patients actually did on continuation medication, the secondary analyses provided an estimate of how those patients would have done if they all had been compliant.

Ease of dissemination. Some treatments are easier to learn than others and therefore easier to disseminate. Classical psychoanalysis required a training analysis of many years' duration, and even those aiming to practice psychoanalytically oriented psychotherapy were encouraged to pursue personal therapy as part of the training process. It is likely that such requirements contributed to the gradual decline in the number of such practitioners across recent decades (Norcross, Karpiak, & Santoro, 2005). No intervention has grown so rapidly over that same period as cognitive-behavior therapy, whereas the number of practitioners professing an allegiance to more purely behavioral therapies has stayed relatively constant over that same interval. That may be subject to change in years to come as the so-called third-wave behavioral approaches gain greater credibility in the treatment community, fueled in part by the perception that it is easier to learn to do simple behavioral interventions than to also learn more complicated cognitive strategies. From a methodological perspective, it is important to report just what was required to implement the

treatment(s) in a given study, including who the therapists were, how they were selected, how much training was required, and what kind of supervision was provided.

Cost-effectiveness. Consideration of the costs of treatment should be conceptually distinct from the scientific evidence for its effectiveness, but costs need to be considered nonetheless (APA, 2002). Costs include the expense to the patient and the health care professional as well as the cost of any technology or equipment involved in the intervention. Clearly, those interventions that produce the same outcome at a lower cost than others are to be preferred. It becomes more difficult to decide on the appropriate course of action when the most expensive treatments are also the most efficacious. Health care economists have developed ways to quantify the costs of leaving problems unresolved (e.g., how much is a depression-free day worth?), and it is often possible to evaluate the relative costs of an intervention against the value it provides. Conducting a sophisticated cost-effectiveness analysis requires collecting information not just on the direct costs of treatment but also on the indirect costs incurred, such as time lost from work and child care and transportation expenses. Moreover, given that costs are often not normally distributed (a single hospitalization can be extremely expensive), the sample sizes required to conduct a sophisticated cost-benefit analysis are often exponentially larger than those required to detect a treatment effect. Nonetheless, information regarding the differential costs and benefits of different treatments can be valuable in evaluating the relative merits of different treatments.

Relative cost-effectiveness also can be influenced by the health care system in which it is embedded. In the United States, health care organizations often have little incentive to provide preventive interventions even when they are cost-effective over the long run because job change happens with sufficient frequency that third-party payers rarely profit from the long-term benefits of preventive care (Stricker et al., 1999). Conversely, in Great Britain, with its single-payer system, the National Health Service has invested £130 million to train therapists to provide

the cognitive and behavioral interventions shown to have enduring effects that make them more cost-effective than long-term medication treatment (Clark et al., 2009). This means that cost-effectiveness is not absolute and must be considered in the context of the larger economic system in which it is embedded.

CONCLUSION

Methodology for intervention research necessarily cuts across many of the topics of other chapters in these volumes, and we have only touched on complex issues that deserve prolonged discussion. We refer the reader to other chapters in these volumes for additional information and also to excellent texts such as Shadish et al. (2002) and Kazdin (2003) on research design and MacKinnon (2008) on mediation. In addition, we believe the reviewer will find it useful to refer to the CONSORT statements on non-inferiority and equivalence trials (Piaggio et al., 2006) and randomized controlled trials (Altman et al., 2001; Boutron, Moher, Altman, Schulz, & Ravaud, 2008), the TREND statement on quasi-experiments (Des Jarlais, Lyles, & Crepaz, 2004), and the APA's paper on journal article reporting standards (APA, Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). Although these statements are designed to encourage uniform reporting of critical design features, being reminded of what these features are when planning the research rather than after the fact facilitates the conduct of sound research.

References

- Addis, M. E., Hatgis, C., Cardemil, E., Jacob, K., Krasnow, A. D., & Mansfield, A. (2006). Effectiveness of cognitive-behavioral treatment for panic disorder versus treatment as usual in a managed care setting: 2-year follow-up. *Journal of Consulting and Clinical Psychology, 74*, 377–385. doi:10.1037/0022-006-X.74.2.377
- Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., . . . Lang, T. (2001). The Revised CONSORT Statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134*, 663–694.

- American Psychological Association. (2002). Criteria for evaluating treatment guidelines. *American Psychologist*, 57, 1052–1059. doi:10.1037/0003-066X.57.12.1052
- American Psychological Association, Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839–851. doi:10.1037/0003-066X.63.9.839
- American Psychological Association, Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, 61, 271–285. doi:10.1037/0003-066X.61.4.271
- Baron, R. M., & Kenny, D. A. (1986). The moderator/mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182. doi:10.1037/0022-3514.51.6.1173
- 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. doi:10.1037/0022-006X.71.6.973
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research, and Practice*, 16, 252–260. doi:10.1037/h0085885
- Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., & Ravaud, P. (2008). Extending the CONSORT Statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration. *Annals of Internal Medicine*, 148, 295–309.
- Brewin, C. R., & Bradley, C. (1989). Patient preferences and randomised clinical trials. *British Medical Journal*, 299, 313–315. doi:10.1136/bmj.299.6694.313
- Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101, 147–158. doi:10.1037/0033-2909.101.1.147
- Burns, D. D., & Spangler, D. L. (2000). Does psychotherapy homework lead to improvements in depression in cognitive-behavioral therapy or does improvement lead to increased homework compliance? *Journal of Consulting and Clinical Psychology*, 68, 46–56. doi:10.1037/0022-006X.68.1.46
- Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). CONSORT Statement: Extension to cluster randomised trials. *BMJ (Clinical Research Ed.)*, 328, 702–708. doi:10.1136/bmj.328.7441.702
- Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, 66, 7–18. doi:10.1037/0022-006X.66.1.7
- Chevron, E. S., & Rounsaville, B. J. (1983). Evaluating the clinical skills of psychotherapists: A comparison of techniques. *Archives of General Psychiatry*, 40, 1129–1132.
- Chlebowski, R. T., Hendrix, S. L., Langer, R. D., Stefanick, M. L., Gass, M., Lane, D., . . . McTiernan, A. (2003). Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women. *Journal of the American Medical Association*, 289, 3243–3253. doi:10.1001/jama.289.24.3243
- Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R., & Wright, B. (2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy*, 47, 910–920. doi:10.1016/j.brat.2009.07.010
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. doi:10.1177/001316446002000104
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Crits-Christoph, P., Baranackie, K., Kurcias, J. S., Beck, A. T., Carroll, K., Perry, K., . . . Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, 1, 81–91. doi:10.1080/10503309112331335511
- Crits-Christoph, P., Tu, X., & Gallop, R. (2003). Therapists as fixed versus random effects—some statistical and conceptual issues: A comment on Siemer and Joormann (2003). *Psychological Methods*, 8, 518–523. doi:10.1037/1082-989X.8.4.518
- Cuijpers, P., van Stratten, A., van Oppen, P., & Andersson, A. (2008). Are psychological and pharmacological interventions equally effective in the treatment of adult depressive disorders? A meta-analysis of comparative studies. *Journal of Clinical Psychiatry*, 69, 1675–1685. doi:10.4088/JCP.v69n1102
- de Jong, K., Moerbeek, M., & van der Leeden, R. (2010). A prior power analysis in longitudinal three-level multilevel models: An example with therapist effects. *Psychotherapy Research*, 20, 273–284. doi:10.1080/10503300903376320
- DeRubeis, R. J., & Feeley, M. (1990). Determinants of change in cognitive therapy for depression. *Cognitive Therapy and Research*, 14, 469–482. doi:10.1007/BF01172968
- DeRubeis, R. J., Gelfand, L. A., Tang, T. Z., & Simons, A. D. (1999). Medications versus cognitive behavior therapy for severely depressed outpatients: Mega-analysis of four randomized comparisons. *American Journal of Psychiatry*, 156, 1007–1013.
- DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., . . . Gallop, R. (2005). Cognitive therapy vs. medications in the

- treatment of moderate to severe depression. *Archives of General Psychiatry*, 62, 409–416. doi:10.1001/archpsyc.62.4.409
- Des Jarlais, D. C., Lyles, C., & Crepaz, N. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND Statement. *American Journal of Public Health*, 94, 361–366. doi:10.2105/AJPH.94.3.361
- Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58, 403–417. doi:10.1093/biomet/58.3.403
- Elkin, I., Gibbons, R. D., Shea, M. T., Sotsky, S. M., Watkins, J. T., Pilkonis, P. A., & Hedeker, D. (1995). Initial severity and differential treatment outcome in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology*, 63, 841–847. doi:10.1037/0022-006X.63.5.841
- Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., . . . Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry*, 46, 971–982.
- Feeley, M., DeRubeis, R. J., & Gelfand, L. A. (1999). The temporal relation of adherence and alliance to symptom change in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, 67, 578–582. doi:10.1037/0022-006X.67.4.578
- Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *JAMA*, 303, 47–53. doi:10.1001/jama.2009.1943
- Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., . . . Weissman, M. M. (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder: Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry*, 48, 851–855.
- Goldfried, M. R. (1980). Toward the delineation of therapeutic change principles. *American Psychologist*, 35, 991–999. doi:10.1037/0003-066X.35.11.991
- Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64–78. doi:10.1037/1082-989X.2.1.64
- Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology*, 60, 73–79. doi:10.1037/0022-006X.60.1.73
- Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ (Clinical Research Ed.)*, 319, 670–674.
- Hollon, S. D. (1996). The efficacy and effectiveness of psychotherapy relative to medications. *American Psychologist*, 51, 1025–1030. doi:10.1037/0003-066X.51.10.1025
- Hollon, S. D., & Beck, A. T. (1978). Psychotherapy and drug therapy: Comparisons and combinations. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (2nd ed., pp. 437–490). New York, NY: Wiley.
- Hollon, S. D., DeRubeis, R. J., Evans, M. D., Wiemer, M. J., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1992). Cognitive therapy, pharmacotherapy and combined cognitive-pharmacotherapy in the treatment of depression. *Archives of General Psychiatry*, 49, 774–781.
- Hollon, S. D., DeRubeis, R. J., Shelton, R. C., Amsterdam, J. D., Salomon, R. M., O'Reardon, J. P., . . . Gallop, R. (2005). Prevention of relapse following cognitive therapy versus medications in moderate to severe depression. *Archives of General Psychiatry*, 62, 417–422. doi:10.1001/archpsyc.62.4.417
- Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, 38, 139–149.
- Jacobson, N. S., & Christensen, A. (1996). Studying the effectiveness of psychotherapy: How well can clinical trials do the job? *American Psychologist*, 51, 1031–1039. doi:10.1037/0003-066X.51.10.1031
- Jacobson, N. S., & Hollon, S. D. (1996). Cognitive behavior therapy vs. pharmacotherapy: Now that the jury's returned its verdict, it's time to present the rest of the evidence. *Journal of Consulting and Clinical Psychology*, 64, 74–80. doi:10.1037/0022-006X.64.1.74
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19. doi:10.1037/0022-006X.59.1.12
- Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology*, 150, 327–333.
- Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston, MA: Allyn and Bacon.
- Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, 57, 138–147. doi:10.1037/0022-006X.57.1.138

- King, M., Nazareth, I., Lampe, F., Bower, P., Chandler, M., Morou, M., . . . Lai, R. (2005). Impact of participant and physician intervention preferences on randomized trials: A systematic review. *JAMA*, 293, 1089–1099. doi:10.1001/jama.293.9.1089
- Klein, D. N., Schwartz, J. E., Santiago, N. J., Vivian, D., Vocisano, C., Castonguay, L. G., . . . Keller, M. B. (2003). Therapeutic alliance in depression treatment: Controlling for prior change and patient characteristics. *Journal of Consulting and Clinical Psychology*, 71, 997–1006. doi:10.1037/0022-006X.71.6.997
- Kraemer, H. C. (2000). Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin*, 26, 533–541.
- Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, 59, 990–996. doi:10.1016/j.biopsych.2005.09.014
- Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, 160, 1566–1577. doi:10.1176/appi.ajp.160.9.1566
- Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, 59, 877–883. doi:10.1001/archpsyc.59.10.877
- Lavori, P. W. (1992). Clinical trials in psychiatry: Should protocol deviation censor patient data? *Neuropsychopharmacology*, 6, 39–48.
- Lavori, P. W., Rush, A. J., Wisniewski, S. R., Alpert, J., Fava, M., Kupfer, D. J., . . . Trivedi, M. (2001). Strengthening clinical effectiveness trials: Equipoise-stratified randomization. *Biological Psychiatry*, 50, 792–801. doi:10.1016/S0006-3223(01)01223-9
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Luborsky, L., & DeRubeis, R. J. (1984). The use of psychotherapy treatment manuals: A small revolution in psychotherapy research style. *Clinical Psychology Review*, 4, 5–14. doi:10.1016/0272-7358(84)90034-5
- Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., . . . Schweizer, E. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, 6, 95–106. doi:10.1093/clipsy/6.1.95
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Erlbaum.
- Manson, J. E., Hsia, J., Johnson, K. C., Rossouw, J. E., Assaf, A. R., Lasser, N. L., . . . Cushman, M. (2003). Estrogen plus progestin and the risk of coronary heart disease. *New England Journal of Medicine*, 349, 523–534. doi:10.1056/NEJMoa030808
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects: An exercise in adversarial collaboration. *Psychological Science*, 12, 269–275. doi:10.1111/1467-9280.00350
- Miller, W. R., Sorensen, J. L., Selzer, J. A., & Brigham, G. S. (2006). Disseminating evidence-based practices in substance abuse treatment: A review with suggestions. *Journal of Substance Abuse Treatment*, 31, 25–39. doi:10.1016/j.jsat.2006.03.005
- Miranda, J., Chung, J. Y., Green, B. L., Krupnick, J., Siddique, J., Revicki, D. A., & Belin, T. (2003). Treating depression in predominantly low-income young minority women. A randomized controlled trial. *JAMA*, 290, 57–65. doi:10.1001/jama.290.1.57
- Mohr, D. C., Spring, B., Freedland, K. E., Beckner, V., Arean, P., Hollon, S. D., . . . Kaplan, R. (2009). The selection and design of control conditions for randomized controlled trials of psychological interventions. *Psychotherapy and Psychosomatics*, 78, 275–284. doi:10.1159/000228248
- Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, 15, 1044–1045. doi:10.1093/beheco/arh107
- Nathan, P. E., Stuart, S. P., & Dolan, S. L. (2000). Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? *Psychological Bulletin*, 126, 964–981. doi:10.1037/0033-2909.126.6.964
- Norcross, J. C., Karpiak, C. P., & Santoro, S. O. (2005). Clinical psychologists across the years: The Division of Clinical Psychology from 1960 to 2003. *Journal of Clinical Psychology*, 61, 1467–1483. doi:10.1002/jclp.20135
- Olkin, I. (1995). Meta-analysis: Reconciling the results of independent studies. *Statistics in Medicine*, 14, 457–472. doi:10.1002/sim.4780140507
- Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., & Evans, S. J. W. (2006). Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT Statement. *Journal of the American Medical Association*, 295, 1152–1160. doi:10.1001/jama.295.10.1152
- Project MATCH Research Group. (1997). Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes. *Journal of Studies on Alcohol*, 58, 7–29.
- Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York, NY: McGraw-Hill.

- Rush, A. J., Beck, A. T., Kovacs, M., & Hollon, S. D. (1977). Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research*, 1, 17–37. doi:10.1007/BF01173502
- Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The *Consumer Reports* study. *American Psychologist*, 50, 965–974. doi:10.1037/0003-066X.50.12.965
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.
- Shao, J., & Zhong, B. (2003). Last observation carry-forward and last observation analysis. *Statistics in Medicine*, 22, 2429–2441. doi:10.1002/sim.1519
- Shaw, B. F. (1984). Specification of the training and evaluation of cognitive therapists for outcome studies. In J. Williams & R. Spitzer (Eds.), *Psychotherapy research: Where are we and where should we go?* (pp. 173–188). New York, NY: Guilford Press.
- Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobson, K. S., . . . Imber, S. D. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, 67, 837–846. doi:10.1037/0022-006X.67.6.837
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428. doi:10.1037/0033-2909.86.2.420
- Sloane, R. B., Staples, F. R., Cristol, A. H., Yorkson, N. J., & Whipple, K. (1975). *Psychotherapy versus behavior therapy*. Cambridge, MA: Harvard University Press.
- Smits, J. A. J., Rosenfield, D., McDonald, R., & Telch, M. J. (2006). Cognitive mechanisms of social anxiety reduction: An examination of specificity and temporality. *Journal of Consulting and Clinical Psychology*, 74, 1203–1212. doi:10.1037/0022-006X.74.6.1203
- Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*, 19, 193–198. doi:10.1016/0149-7189(96)00011-0
- Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Brody, P. E. (2003). Are samples in randomized controlled trials of psychotherapy representative of community patients? A new methodology and initial findings. *Journal of Consulting and Clinical Psychology*, 71, 963–972. doi:10.1037/0022-006X.71.6.963
- Stricker, G., Abrahamson, D. J., Bologna, N. C., Hollon, S. D., Robinson, E. A., & Reed, G. M. (1999). Treatment guidelines: The good, the bad, and the ugly. *Psychotherapy: Theory, Research, Practice, Training*, 36, 69–79. doi:10.1037/h0087755
- Striegel-Moore, R. H., Wilson, G. T., DeBar, L., Perrin, N., Lynch, F., Rosselli, F., & Kraemer, H. C. (2010). Cognitive behavioral guided self-help for the treatment of recurrent binge eating. *Journal of Consulting and Clinical Psychology*, 78, 312–321. doi:10.1037/a0018915
- Trivedi, M. H., Rush, A. J., Crismon, M. L., Kashner, T. M., Toprac, M. G., Carmody, T. J., . . . Shon, S. P. (2004). Clinical results for patients with major depressive disorder in the Texas Medication Algorithm Project. *Archives of General Psychiatry*, 61, 669–680. doi:10.1001/archpsyc.61.7.669
- Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, 61, 620–630. doi:10.1037/0022-006X.61.4.620
- Wei, L. J., & Lachin, J. M. (1988). Properties of the urn randomization in clinical trials. *Controlled Clinical Trials*, 9, 345–364. doi:10.1016/0197-2456(88)90048-7
- Weisz, J. R., & Addis, M. E. (2006). The research-practice tango and other choreographic challenges: Using and testing evidence-based psychotherapies in clinical care settings. In C. D. Goodheart, A. E. Kazdin, & R. J. Sternberg (Eds.), *Evidence-based psychotherapy: Where practice and research meet* (pp. 179–206). Washington, DC: American Psychological Association. doi:10.1037/11423-008
- Wennberg, J. E., Barry, M. J., Fowler, F. J., & Mulley, A. (1993). Outcomes research, PORTs, and health care reform. *Annals of the New York Academy of Sciences*, 703, 52–62. doi:10.1111/j.1749-6632.1993.tb26335.x
- Westen, D., Novonty, C. M., & Thompson-Brenner, H. (2004). Empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin*, 130, 631–663. doi:10.1037/0033-2909.130.4.631
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. doi:10.1037/0003-066X.54.8.594