Journal of Consulting and Clinical Psychology
1998, Vol. 66, No. 1, 7–18

0022-006X/98/$3.00

Defining Empirically Supported Therapies

Dianne L. Chambless
University of North Carolina at Chapel Hill

Steven D. Hollon
Vanderbilt University

A scheme is proposed for determining when a psychological treatment for a specific problem or disorder may be considered to be established in efficacy or to be possibly efficacious. The importance of independent replication before a treatment is established in efficacy is emphasized, and a number of factors are elaborated that should be weighed in evaluating whether studies supporting a treatment's efficacy are sound. It is suggested that, in evaluating the benefits of a given treatment, the greatest weight should be given to efficacy trials but that these trials should be followed by research on effectiveness in clinical settings and with various populations and by cost-effectiveness research.

In this special section, the authors of review articles have been asked to evaluate the psychological treatment research literature in their area of expertise to identify empirically supported treatments[1] (ESTs). Briefly, we define ESTs as clearly specified psychological treatments shown to be efficacious in controlled research with a delineated population. Thus, following Kiesler (1966), we suggest that practitioners and researchers will profit from knowing which treatments are effective for which clients or patients. That psychological treatments (undefined) benefit the majority of clients or patients (undefined) is already well established (see, e.g., M. L. Smith & Glass, 1977). Furthermore, our particular emphasis is on the effects of the treatments as independent variables. This is not to deny the importance of other factors such as the therapeutic alliance, as well as client and patient variables that affect the process and outcome of psychological therapies (see, e.g., Beutler, Machado, & Neufeldt, 1994; Garfield, 1994; Orlinsky, Grawe, & Parks, 1994). Rather, these factors are simply not the focus of this special section.

Implementation of the plan for this special section required an operational definition of ESTs. In this article, we provide a structure for evaluation of treatments that the authors have been asked to adopt as a starting point for their reviews. We draw on the foundations provided by the Division 12 (Clinical Psychology) Task Force on Promotion and Dissemination of Psychological Procedures (1995; Chambless et al., 1996) and the American Psychological Association (APA) Task Force on Psychological Intervention Guidelines (1995), but we have made a number of changes. Hence, the responsibility for the particular criteria we describe here is our own. These criteria are summarized in the Appendix and discussed in detail in the section on efficacy.

Evaluators have been asked to consider the following broad issues about ESTs in their area: (a) Has the treatment been shown to be beneficial in controlled research? (b) Is the treatment useful in applied clinical settings and, if so, with what patients and under what circumstances? (c) Is the treatment efficient in the sense of being cost-effective relative to other alternative interventions? These questions are addressed by studies on efficacy (including clinical significance), effectiveness (or clinical utility), and efficiency (or cost-effectiveness).

Efficacy

Overall Research Design

Following the two task forces (APA Task Force on Psychological Intervention Guidelines, 1995; Task Force on Promotion and Dissemination of Psychological Procedures, 1995), we take as our starting point the position that treatment efficacy must be demonstrated in controlled research in which it is reasonable to conclude that benefits observed are due to the effects of the treatment and not to chance or confounding factors such as passage of time, the effects of psychological assessment, or the presence of different types of clients in the various treatment conditions (see Campbell & Stanley, 1963; Kazdin, 1992). In our view, efficacy is best demonstrated in randomized clinical trials (RCTs)—group designs in which patients are randomly assigned to the treatment of interest or one or more comparison conditions—or carefully controlled single case experiments and their group analogues. This approach has not gone unchallenged. Some argue that statistical controls alone can suffice to draw

Dianne L. Chambless, Department of Psychology, University of North Carolina at Chapel Hill; Steven D. Hollon, Department of Psychology, Vanderbilt University.

We thank the members of the Division 12 Task Force on Psychological Interventions, The Division 12 Task Force on Promotion and Dissemination of Psychological Procedures, and the APA Task Force for Psychological Intervention Guidelines for their contributions to the thinking that shaped this article, along with the many individuals who have provided feedback on the task forces' efforts. In particular, we acknowledge Paul Crits-Christoph, Robert Elliott, David Haaga, Varda Shoham, and John Weisz.

Correspondence concerning this article should be addressed to Dianne L. Chambless, Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27599-3270. Electronic mail may be sent to chambles@email.unc.edu.

[1] The term we have elected to use, *empirically supported therapies,* is deliberately different from *empirically validated therapies,* the term used by the American Psychological Association Division 12 Task Force (1995), for two reasons: (a) to make clear that the criteria for the two labels are different and (b) to avoid the unfortunate connotation, to some, of the phrase *empirical validation* (to wit, that the process of validation has been completed, and no further research is needed on a treatment; see Garfield, 1996).

causal inferences (e.g., Seligman, 1995). However, these approaches are so susceptible to model misspecification and inferential error that any conclusion drawn on their basis must be tentative indeed. For this reason, we simply have more confidence in inferences derived from controlled experimentation than those derived from purely correlational analyses, no matter how complex.

As is the case in research in general, replication is critical, particularly replication by an independent investigatory team. The requirement of replication helps to protect the field from drawing erroneous conclusions based on one aberrant finding. Replication by an independent team of investigators also provides some protection against investigator bias or reliance on findings that prove unique to a particular setting or group of therapists. Thus, only when a treatment has been found efficacious in at least two studies by independent research teams do we consider its efficacy to have been established and label it an *efficacious* treatment. If there is only one study supporting a treatment's efficacy, or if all of the research has been conducted by one team, we consider the findings promising but would label such treatments as *possibly efficacious,* pending replication.

Furthermore, we specify that the efficacy research must have been conducted with methods that are adequately sound to justify reasonable confidence in the data. No one definition of sound methodology suffices for all areas of psychological treatment research. Nonetheless, we lay out some of the basic issues that evaluators need to consider after a brief description of the overall designs that might be used in treatment research acceptable for our purposes. We begin our discussion with a focus on group designs because they are more common; we comment later on specific issues for single-case experiments.

Comparisons with no treatment. The fundamental question clients presenting for treatment are likely to pose is, "Does the treatment you propose for me actually work?" This question is addressed by the comparison of treatment with some type of minimal or no-treatment condition (e.g., waiting list or assessment-only control) in which clients receive the assessment procedures but no significant amount of active treatment. If, in two or more studies conducted by different research teams, treatment proves more beneficial than no treatment and the findings are not contradicted by others (see later section on resolution of conflicting data), we consider the treatment to be efficacious. Obviously, the more replications that have been conducted and the more different settings in which these replications have been carried out, the more confidence one has in the findings. Thus, we set here a minimum threshold rather than an optimal one.

In this regard, our criteria differ from those originally proposed by the Division 12 task force (1995) in that we do not require evidence of specificity to consider a treatment efficacious. Although evidence that a given treatment is superior to a pill or psychological placebo or to another treatment informs inferences about causal agency, we suggest that it is not necessary to demonstrate such specificity for a treatment to be said to have a beneficial effect (cf. Parloff, 1986). Simply put, if a treatment works, for whatever reason, and if this effect can be replicated by multiple independent groups, then the treatment is likely to be of value clinically, and a good case can be made for its use.[2]

Comparisons with other treatments or with placebo. It is also important to know whether the mechanisms that underlie an observed effect go beyond the simple consequences of receiving attention from an interested person or the expectation of change. For this reason, treatments found to be superior to conditions that control for such nonspecific processes or to another bona fide treatment are even more highly prized and said to be *efficacious and specific* in their mechanisms of action. Such findings have implications for theory, because they increase confidence in the specific explanatory model on which the treatment is based, but also for practice, because they suggest that particular kinds of training and experience may be necessary to produce the desired effect.

Comparisons with other rival interventions can provide the most stringent tests of all, because they not only control for processes independent of treatment and common to all treatments but may also involve tests between competing specific mechanisms. Thus, such studies sit atop a hierarchy of increasing competitive difficulty and should be at least as informative about causal agency as comparisons with no-treatment conditions or even nonspecific controls. Moreover, such comparisons can provide explicit information regarding the relative benefits of competing interventions. For this reason, treatments that are found to be superior to other rival alternative interventions are more highly valued still.

For ethical reasons, some investigators prefer to compare treatments of unproven benefit with established interventions for the purpose of demonstrating equivalence. Although this design allows the investigator to avoid assigning some clients to conditions he or she believes will be inferior, the resultant data are inherently difficult to interpret. One concern is that the evaluator needs to be confident that the established comparison treatment was implemented well in the particular study; otherwise, the unproven treatment will erroneously be deemed efficacious by comparison with a degraded version of the more established treatment (Klein, 1996). Another common problem is low statistical power. Rarely do psychological treatment researchers have a sufficient number of clients in their study to detect medium differences among treatments with statistical tests of significance. For example, to have the conventional 80% power (Cohen, 1988) for a significance test of a medium difference between two treatment groups, an investigator needs approximately 50 clients per condition, a very expensive proposition indeed. In contrast, in their meta-analysis of psychological treatment outcome research, Kazdin and Bass (1989) found that the median sample size per treatment condition was 12! For these and other reasons, equivalence is always easier to interpret if it occurs in the context of a design that compares the estab-

[2] We recognize that drugs are typically required to demonstrate specificity before they can be brought to market; that is, they must be shown to have a pharmacological effect that transcends the benefits provided by simply believing that one is taking an active medication. We make no corresponding requirement for psychological treatments. If, as some theorists suggest, it were ultimately proved that psychological treatments work solely through shared relational mechanisms (so-called nonspecific effects), it would still be a legitimate enterprise as long as it provided benefit. This represents something of a double standard but one that reflects the twin notions that drugs are particularly likely to cause noxious side effects and that it is not ethical to deceive people about whether they are taking a pharmacologically active medication.

lished efficacious treatment with the kinds of control conditions against which it was initially established.

Whatever qualms we have about interpreting a treatment as efficacious on the basis of null results, we recognize that much of the psychological treatment research literature falls in this category. Rogers, Howard, and Vessey (1993) have recommended a methodology for determining when two groups are statistically equivalent. This method could (and should) be research. Unfortunately, this procedure has yet to find its way into the treatment outcome literature. Rather than ignore most of the comparative psychological treatment literature, we suggest an intermediate step: When, in an otherwise sound study, (a) investigators have a sample size of 25–30 per condition (thus allowing a reasonably stable estimate of the effects of treatment), (b) the unproven treatment is not significantly inferior to the established efficacious treatment on tests of significance, and (c) the pattern of the data indicates no trends for the established efficacious treatment to be superior, the treatments may be considered equivalent in efficacy in this study.

Combination treatments. Finally, the special case of combination treatments needs to be addressed. A typical study of combination treatments involves comparison of a multiple component treatment and one or more of its individual parts (e.g., medical intervention alone for chronic pain vs. medical intervention plus psychological treatment). In keeping with our greater emphasis on replication, we do not think that it is necessary to control for the additional attention and support typically provided in a multiple component treatment to conclude that such combinations are possibly efficacious or established as adjuncts to medication. However, such controls would be necessary to establish the specificity of that effect.

A last concern about combination treatment research involves contrasts among psychological treatment, pharmacotherapy, and their combination. A common design has four cells: drug, placebo, psychological treatment plus drug, and psychological treatment plus placebo. In cases in which psychological treatment plus placebo is better than placebo alone, investigators often indicate that they have demonstrated the efficacy of the psychological intervention used on its own. Such an assumption is unwarranted, because it is possible that the effects of psychological treatment plus placebo are not simply additive but interactive. That is, psychological treatment may be potent only in combination with the placebo. Consider, for example, the hypothetical case of a man with social phobia who agrees to test himself in previously avoided situations because the placebo gives him confidence that he will not be too anxious to talk to a woman he is interested in dating. In such a case, exposure instructions would be beneficial, but only in the context of the morale boost from the placebo. Thus, in the absence of a fifth cell in which patients receive only psychological treatment, confident conclusions about the efficacy of the psychological intervention are precluded (for a further discussion of these issues, see Hollon & DeRubeis, 1981).

Sample Description

In light of the great heterogeneity of problems for which psychologists provide treatment, we believe that, if psychological treatment outcome research is to be informative, researchers must have clearly defined the population for which the treatment was designed and tested. Thus, we do not ask whether a treatment is efficacious; rather, we ask whether it is efficacious for a specific problem or population.

In much recent research, the sample is described in terms of a diagnostic system such as the *Diagnostic and Statistical Manual of Mental Disorders* (third edition revised; American Psychiatric Association, 1987). This approach has a number of benefits in that it ties treatment research to a large body of descriptive psychopathology literature based on the same definitions, and there are standardized diagnostic interviews that permit reliable diagnoses to be made. Moreover, many clinicians are familiar with this nomenclature because they use this system for their own diagnostic work or billing for third-party payers. Diagnostic labels, however, are not the only reliable method for identifying a population for research. Alternatives include cutoff scores on reliable and valid questionnaires or interviews identifying the problem or focus of interest. Consider, for example, a marital discord prevention program designed for couples planning to marry in the subsequent 6 months. Being engaged is hardly a disease state, but it can be reliably identified. In addition to a description of the presenting problem, investigators should describe other characteristics of the population that might affect the generalizability of their findings (e.g., major comorbid conditions included or excluded, age range, and socioeconomic status).

When the sample is described in terms of a diagnostic system, it is highly desirable that the investigators use a structured diagnostic interview to assign diagnoses and that they demonstrate that these diagnoses were reliably made in their study. In many older studies, this step may have been omitted. Whether such studies are useful depends on the problem in question. Some diagnoses are easily made, and their definition has changed little over the years (e.g., obsessive–compulsive disorder), whereas others are more difficult and unreliable (e.g., generalized anxiety disorder). The importance of standardized diagnostic procedures increases along with the difficulty of the diagnosis.

Outcome Assessment

Selection of instruments. Given that we have specified that we want to know whether a treatment is efficacious for a particular problem, it follows that outcome assessment tools need to tap the significant dimensions of that problem. These tools should have demonstrated reliability and validity in previous research; reliability of interviewer-rated measures should be demonstrated; and interviewers should be uninformed as to the treatment group assignment of clients they assess. It is desirable, although not absolutely required, that multiple methods of assessment be used, and it is particularly desirable that researchers not rely solely on self-report.

In some areas of research, investigators rely on indexes of such high face validity that consideration of formal psychometric properties beyond reliability would add little. Examples include pounds lost and reductions in the frequency of arrests or days in the hospital. We do not insist on prior validity studies in such cases. However, for most measures, some explicit attention needs to be paid to the construct validity of the measures selected (Cronbach & Meehl, 1955).

Similarly, it is desirable that investigators go beyond assessment of symptoms and examine the effects of treatment on more

general measures of functioning and quality of life. Not all life concerns can be reduced to specific signs and symptoms, and treatments may differ with respect to the breadth of their effects. Moreover, it is important that they consider whether there are negative effects of treatment. In practice, few studies as yet have included such measures, but we expect to see more such data emerging in future research.

Follow-up. Information regarding the long-term effects of treatment is highly desirable but difficult to come by. At the least, it is important to know whether treatment has an enduring effect and whether different treatments differ with respect to their stability over time. Psychological treatments have long been presumed to have more stable effects than pharmacotherapy, either because they redress underlying propensities that contribute to risk or because patients acquire stable skills that enable them to cope better with those underlying vulnerabilities. There is growing evidence that this may be the case for at least the cognitive and cognitive–behavioral interventions (Hollon & Beck, 1994) and tantalizing evidence that interpersonal psychotherapy may have a delayed effect on social functioning and related symptoms that may not emerge until well after treatment has been completed (Fairburn, Jones, Peveler, Hope, & O'Connor, 1993; Weissman & Markowitz, 1994).

Nonetheless, follow-up studies are hard to conduct and difficult to interpret. Patients are typically free to pursue additional treatment if they so desire, and they may do so for a number of reasons. They are particularly likely to go back into treatment if their symptoms return, but they may also return to treatment for other reasons. Moreover, return to treatment may either prevent an incipient symptomatic relapse or resolve a problem if it has already started. Thus, it is not clear whether return to treatment in the absence of a documentable symptomatic event should be considered an index of underlying risk.

The situation can be further complicated if both treatment and symptoms status are not assessed in an ongoing fashion over time. Without longitudinal assessment, clients' status at any one point can be misleading. For example, Nicholson and Berman (1983) once questioned whether follow-up studies were really necessary because the existing studies so rarely produced evidence of differential stability of effects. What the authors failed to recognize is that differential return to treatment often masks the emergence of differences in the stability of response. Because patients are particularly likely to pursue additional treatment if symptoms begin to return, and because subsequent treatment may redress those emerging symptoms, simple cross-sectional assessments that do not take into account the sequence of events over time may fail to detect real prophylactic effects. For example, depressed patients treated to remission pharmacologically are more likely both to relapse and to seek subsequent treatment after initial termination than are patients treated to remission with cognitive therapy, the net effects of which are likely to cancel each other out when assessed only cross sectionally at periodic intervals (Hollon, Shelton, & Loosen, 1991).

Strategies that monitor the occurrence of symptomatic events across time (e.g., survival analyses) provide some protection against the potential masking effects of differential treatment return (Greenhouse, Stangl, & Bromberg, 1989). Nonetheless, they are not without their own difficulties, because they tend to be low in power and do little to resolve the ambiguity surrounding premature return to treatment (i.e., return to treatment in the absence of a documentable symptomatic relapse). Investigators can conduct multiple analyses in which premature return is sometimes ignored, sometimes incorporated into the definition of relapse, and sometimes treated as a censoring variable, thereby removing the participant from the pool at risk for relapse (Evans et al., 1992).

Another major problem with follow-up designs is that they are particularly susceptible to bias resulting from differential retention. In most naturalistic follow-ups, patients not only must complete treatment but must show improvement to be retained in the sample. If one treatment is more effective than another in keeping high-risk patients in treatment or in eliciting their response, then acute treatment can act as a differential sieve that undermines the effects of initial randomization (Klein, 1996). For example, several studies have suggested that patients treated to remission with cognitive therapy are about half as likely to relapse after treatment termination as patients who respond to pharmacotherapy (Hollon et al., 1991). However, only about half of the patients initially assigned to treatment both complete and respond to either modality. If high-risk patients need medications to respond and are better able to tolerate their side effects, then high-risk patients would be systematically screened out of cognitive therapy because they did not show improvement. Concomitantly, low-risk patients would be systematically screened out of pharmacotherapy because they could not tolerate the medication. The consequence would be that underlying risk would no longer be randomly distributed between the treatment conditions, and cognitive therapy's apparent prophylactic effect could be nothing more than an artifact of its lower capacity to retain high-risk patients (Klein, 1996).

There is no simple resolution to this problem. Lavori (1992) has recommended retaining all patients initially assigned to treatment in the final analysis (intention-to-treat analysis), but it is not clear how to categorize patients who either drop out of treatment or initially fail to respond. Censoring their data from the beginning of the follow-up (i.e., ignoring information about their subsequent course) is functionally equivalent to excluding them from the analysis, whereas treating them as if they exhibited a symptomatic event at the beginning of the follow-up period confounds conceptually different negative outcomes. Neither attrition nor nonresponse is functionally equivalent to relapse or recurrence. They may all be undesirable outcomes, but it is likely that they have different causes and consequences. Nonetheless, we think it is better to collect these data (and to exercise caution in their interpretation) than to not collect them at all.

Finally, it is not clear how long naturalistic follow-ups ought to be maintained. Any information is likely to have some value, but the utility of continuing data collection is likely to reach a point of diminishing returns. There is no uniform answer to the question of how long posttreatment follow-ups ought to be maintained. We suggest that those disorders that tend to follow a more variable course require longer follow-ups than those that tend to be more stable over time. Similarly, the magnitude of the potential prophylactic effect is also a consideration; strong effects are likely to require less time to be detected than weak ones, particularly if they are stable over time. Thus, the length of the follow-up required is likely to depend on the natural course of the disorder in question and the strength and stability of the treatment effect that is worthy of detection.

Clinical significance. Of great importance, but often ignored, is an assessment of the clinical significance of treatment response. If a treatment is to be useful for practitioners, it is not enough for treatment effects to be statistically significant; they also need to be large enough to be clinically meaningful. Theoretically, it is possible for a small effect to be so robust that it exceeds anything that could be explained by chance alone but still not be of sufficient magnitude to be of value clinically, particularly if the sample studied is large.

There are a number of ways to assess the clinical significance of an effect. Jacobson and colleagues have developed an approach to assessing clinical significance that defines reliable change in terms of the error of measurement and defines meaningful change in terms of the intersection between functional and dysfunctional populations; clinical significance is indicated by the proportion of patients who meet both criteria (Jacobson & Truax, 1991). Other investigators have compared individual outcomes against a normative standard. Procedures for such comparisons have been described by Kendall and Grove (1988). For some populations a return to normalcy is a reasonable goal, whereas for others (e.g., patients with chronic schizophrenia) it is not. Here, clinical significance might be determined on the basis of attainment of some other societally or personally important goal (e.g., the number of patients able to live in a group home with minimal supervision). The evaluators have been urged to report available information on the clinical significance of findings. However, we have not raised attainment of a particular level of clinical significance to the status of a criterion for ESTs for two reasons. First, authors have only recently begun to provide such data. Second, the issue of how much change to require for a given problem is quite complex (e.g., If spouses who have attended marital therapy do not divorce, is this necessarily a good outcome?), and such a decision is not readily generalized across presenting problems.

Treatment Implementation

Treatment manuals. Psychological treatment research is not informative to the field if one does not know what treatment was tested; nor can researchers replicate an undefined treatment intervention. For this reason, research projects for which a treatment manual was not written and followed are of limited utility in terms of assessment of treatment efficacy.[3] The exception is a treatment intervention that is relatively simple and is adequately specified in the procedure section of the journal article testing its efficacy.

Treatments manuals are, at base, a cogent and extensive description of the treatment approach therapists are to follow. Depending on the type of psychological treatment to be tested, they may contain careful session-by-session outlines of interventions, or they may describe broad principles and phases of treatment with examples of interventions consistent with these notions. It is unlikely that these manuals will be sufficient in themselves. Usually they will need to be supplemented by additional training and supervision, but they should provide a clear and explicit description of the kinds of techniques and strategies that constitute the intervention.

Therapist training and monitoring. It is important to ascertain whether the treatments provided in a given study were implemented in an adequate fashion. A particular intervention may fail to impress not because it lacks efficacy but because it was poorly implemented as a result of inadequate training or supervision. This is particularly likely to be the case when therapists are new to a modality or have an allegiance to another approach. For example, despite the evident care that went into training, the recent National Institute of Mental Health Treatment of Depression Collaborative Research Program has been criticized for failing to provide sufficient supervision to ensure that recently trained cognitive therapists (several of whom had prior allegiances to other approaches) could and would implement that modality in a competent fashion (Jacobson & Hollon, 1996a, 1996b).

Our concern with therapist training flies in the face of considerable research suggesting that therapist experience contributes little to outcome (Christensen & Jacobson, 1994). We are skeptical of this literature and the way it is being interpreted. Many of the studies looked at credentials rather than competence or at years of experience in conducting any type of psychological treatment versus training in and experience with a specific EST. We would not expect that number of years of experience in administering a minimally beneficial treatment would matter; when there are specific efficacious interventions to learn, however, we do expect training to matter. Moreover, even the prototypic studies in this literature fail to stand up to careful scrutiny. For example, Strupp and Hadley's (1979) oft-cited finding that kindly college professors performed about as well as experienced psychodynamic or experiential therapists provides a case in point: The college professors were highly selected for their skill in working with the minimally troubled undergraduates who made up the bulk of the sample, whereas the trained psychotherapists were not. There is considerable evidence that therapist effects are reduced in controlled clinical trials relative to naturalistic studies of clinical practice (Crits-Christoph et al., 1991), presumably in part as a result of the greater training and supervision provided, as well as initial selection of talented therapists. This is not to say that all types of therapies require trained and experienced practitioners to produce maximum effect, but there is evidence that this is true for at least some types of interventions and for some populations (e.g., Burns & Nolen-Hoeksema, 1992; Lyons & Woods, 1991; Weisz, Weiss, Han, Granger, & Morton, 1995).

The problem is compounded by the fact that there is, at present, only a rudimentary sense of how best to measure quality of implementation. Certainly, any such assessment should be based on actual samples of therapist behavior, but the measurement technologies that will make that possible are only now beginning to be developed. It appears to be easier to measure adherence (whether therapists did what they were supposed to do) than it is to measure competence (how well they did it; Waltz, Addis, Koerner, & Jacobson, 1993). Research supporting the commonsense notion that therapists' competence should be related to treatment outcome is beginning to appear (e.g., Barber, Crits-Christoph, & Luborsky, 1996). However, the use of

[3] Psychotherapy process research may be less hindered by the lack of a treatment manual. For example, one could investigate whether the therapist–client relationship predicts outcome or what therapist behaviors lead to a better working alliance without specifying the nature of the interventions closely. However, our focus in this special section is on the effects of treatment rather than the process.

these measures is fairly recent. Accordingly, we have asked evaluators to consider studies in which checks for adherence or competence were not conducted but to comment on this drawback.

Investigator allegiance. Both quantitative and qualitative reviews have consistently suggested that outcome variability across studies is associated with the preferences and expertise of the respective research teams involved (Luborsky, Singer, & Luborsky, 1975; Robinson, Berman, & Neimeyer, 1990; M. L. Smith & Glass, 1977). That is, any given therapy tends to do better in comparison with other interventions when it is conducted by people who are expert in its use than when it is not. For example, cognitive therapy for panic has fared better (relative to pharmacotherapy) when it has been implemented by knowledgeable experts (Clark et al., 1994) than when it has not (Black, Wesner, Bowers, & Gabel, 1993). Conversely, differences favoring cognitive therapy over drugs in the treatment of depression found in early comparisons have not been evident in more recent studies that have done a better job of implementing pharmacotherapy in an adequate fashion (Hollon et al., 1991).

Our sense is that this has more to do with honest differences in the capacity of any given research group to adequately implement multiple interventions than it does with any more malignant effort to bias the results. It seems to us that the best way to deal with this issue is not to try to eliminate allegiances but to balance them. We are particularly impressed by studies in which each modality is overseen by knowledgeable experts committed to ensuring its competent implementation. To date, the number of such studies is limited. Accordingly, evaluators need to be alert to the potential influence of allegiance effects as they evaluate the findings in a given area of research. Inferences regarding treatment efficacy can be framed only in the context of how the therapy was delivered and by whom.

Data Analysis

Although the peer review process often ensures that authors will have analyzed and interpreted their outcome data appropriately, we find this is not always the case. For this reason, evaluators have been asked to make their own judgments of the results of each study purporting to demonstrate treatment efficacy.

We have previously mentioned the problem of low statistical power, wherein investigators may decide two treatments are equally efficacious on the basis of nonsignificant statistical tests, even though their sample size is too small to detect differences that are clinically important. The following are other typical errors in analysis and interpretation we have observed.

1. The authors conduct many tests, find one that is significant and advantageous to their favored treatment, and conclude they have demonstrated its superiority (Type I error). This would be credible only if there were a convincing and a priori rationale that this one measure is the critical outcome measure.

2. Rather than relying on between-groups comparisons when examining the efficacy of their favored treatment versus a waiting list control or other treatment procedures, the authors (a) report the uncontrolled pretest–posttest comparisons; (b) note that the favored treatment showed significant change from baseline, whereas the control condition did not; and (c) conclude that, therefore, their treatment has been shown to be superior to the control condition. This argument is specious because the favored treatment may have barely met the criterion for significance and the control condition have barely missed it, all in the absence of even a trend for the groups to differ significantly. Furthermore, we have already described the reasons that uncontrolled pretest–posttest comparisons do not provide interpretable data.

3. The treatments have different rates of refusal or dropout, but the authors ignore this problem in their data analysis or interpretation. For example, consider the hypothetical example of Treatment A, in which 70% of clients improve, and Treatment B, in which 50% of clients improve. Treatment A would appear to be more efficacious than Treatment B, unless the authors take into account that 50% of the clients who entered Treatment A dropped out, whereas 10% of those who entered Treatment B failed to complete treatment. Thus, of the original group that started Treatment A, 35% completed treatment and improved, whereas, of those who began Treatment B, 45% completed treatment and improved. Whenever there is differential attrition from treatment groups, authors need to conduct such intention-to-treat analyses in which the outcomes for all individuals who were randomized to treatments are examined (see Flick, 1988).

4. The investigators fail to test for therapist or site effects. We have argued that evaluators need to be alert to the possibilities of site effects (which include allegiance of investigator effects, as well as the effects of experience with a given treatment modality) or inadequate therapist training or monitoring. So do investigators. Checks for such effects should be planned as part of the data analyses. Ideally, any given study should involve multiple therapists so as to enhance the generality of the findings, and the investigators should test whether the therapists are differentially effective. Similarly, tests should be made for site effects before pooling across settings in multisite studies. In either case, it is not sufficient to find that these factors fall short of conventional levels of significance; rather, investigators should relax their criterion for statistical significance somewhat to make sure that they do not inadvertently overlook a potentially confounding variable (Crits-Christoph & Mintz, 1991). Because the inclusion of such analyses is relatively recent, we have asked evaluators to consider site and therapist effect analyses as desirable, but not required, for efficacy tests.

Single-Case Experiments

In general, the principles we have described for evaluating group design psychological treatment research also apply to single-case experiments. In addition, there are special issues to consider (see Barlow & Hersen, 1984; Kazdin, 1982).

Establishing a stable baseline. To be able to demonstrate that a treatment has changed a target behavior, single-case experimenters first need to establish a baseline for that behavior over a period of time. The baseline serves as the comparison condition that controls for the effects of assessment and the passage of time and that, depending on the nature of the baseline (no intervention vs. control intervention), may control for expectancy, attention, and the like. Often, during baseline, improvement due to one or more of these factors will be noted. The investigator should implement the treatment of interest only once the baseline of the behavior targeted for change is stable or indicates deterioration for at least the three assessment points

necessary to establish a linear trend. If the behavior examined is variable (e.g., a client with considerable daily fluctuations in severity of depressed mood), longer baselines are needed to establish the pattern of the data before proceeding.

Typical acceptable designs. Because the types of single-case experimental designs are limited only by the creativity of the experimenter, it is impossible to enumerate all possible acceptable designs. However, we mention the most common.

1. ABAB design in which A is the baseline condition and B is the treatment of interest: Once an appropriate baseline is established in the first A period, the pattern of the data must show improvement during the first B period, reversal or leveling of improvement during the second A period, and resumed improvement in the second B period. This design provides a powerful demonstration that the effects of treatment in B are not due to passage of time, changes in the client's life external to the treatment process, or the effects of continued assessment alone. This design can be extended to a similar group design, the equivalent time-samples design (Campbell & Stanley, 1963). According to this design, a single group of participants undergoes multiple periods of the active treatment and of the control condition. When random assignment is used to determine which intervals will contain active treatment and which will contain the control condition, this design allows a high level of confidence in efficacy data collected within a single group of participants.

2. Multiple baseline designs: There are several variations on this theme, including multiple baselines across behaviors, settings, and participants. Note that the last is not literally a single case design, but we include it here because of its close similarity to such designs.

In multiple baseline across behaviors designs, the researcher must identify at least three clinically important behaviors that are relatively independent of one another, that is, behaviors that are not so closely linked that a change in one brings about change in the others without specific intervention for the two behaviors yet to be targeted. After establishing the initial baseline for all behaviors, the experimenter must show that Behavior 1 changes when it is the target of treatment, but Behaviors 2 and 3 remain stable or deteriorate; that Behavior 2 changes when and only when it becomes the focus of treatment, whereas Behavior 3 remains stable or deteriorates; and that, in turn, Behavior 3 changes once it is included as a treatment focus.

Multiple baselines across settings target a similar behavior in several different situations (e.g., home, classroom, and playground), whereas multiple baselines across participants involve 3 different participants rather than three different behaviors. In all cases, the pattern of the data needs to be as described for multiple baselines across settings if the data are to be clearly interpreted. The multiple, linked replications and the timing of change make it highly implausible that assessment effects, the passage of time, or external events in a patient's life are responsible for the pattern of improvement observed.

Defining efficacy on the basis of single-case experiments. Our categorization of the results of single-case experiments, like those of group designs, rests on the number and independence of replications. We consider a treatment to be possibly efficacious if it has proved beneficial to at least 3 participants in research by a single group. Multiple replications (at least three each) by two or more independent research groups are required before we consider a treatment's efficacy as established (each in the absence of conflicting data). If, during the baseline phase (or phases), the client is engaged in an alternative treatment controlling for expectancy and attention as well as assessment effects, we consider the effects to be specific.

Interpretation of results. Researchers usually interpret single-case experiments visually because they are interested in effects that are so striking that they are readily convincing to the naked eye. Comparisons of active treatment and an assessment-only condition often provide such stark contrasts. Comparisons among active treatments are not always so obvious. Moreover, the naked eyes of different judges frequently disagree. Accordingly, statistical procedures have been introduced (e.g., ITSACORR; Crosbie, 1993), but these procedures require many data points and have been used in relatively few studies. Given the potential for investigator bias, evaluators are urged to carefully examine data graphs and draw their own conclusions about the efficacy of the intervention.

Resolution of Conflicting Results

In the previous sections of this article, we have repeatedly mentioned the caveat "in the absence of conflicting results," but, of course, conflicting results are not unusual in psychological research. Evaluators need to consider the entire body of controlled research literature on a treatment when they consider its efficacy. Thus, for example, when we state that the efficacy of a treatment has been established if it has been shown to be better than an assessment-only or waiting list control condition in two studies conducted by two teams, we do not mean that the threshold is crossed whenever two positive studies appear, regardless of the amount of conflicting data available. How are evaluators to resolve discrepancies?

First, they examine the quality of the conflicting research. If the well-designed studies point in one direction and the poorly designed studies point in another, the well-designed studies carry the day. Second, in a group of studies of roughly comparable methodological rigor, evaluators consider whether the preponderance of studies argue for the treatment's efficacy. When the picture is truly mixed, we ask that evaluators be conservative and not include the treatment as possibly efficacious or established in efficacy until such time as the factors leading to differential results can be identified. Third, evaluators are urged to consult meta-analyses while realizing their limitations as well as their benefits. These limitations and benefits bear some discussion.

Meta-analyses can provide useful summaries of large bodies of empirical studies and provide one means to compensate for the limited power of individual studies. However, they can also obscure qualitative differences in treatment execution. For example, Dobson (1989) found that cognitive therapy outperformed drugs in the treatment of depression by a magnitude of about half a standard deviation (a moderate-sized effect that would have real implications for clinical practice) but failed to recognize that the observed effect was largely due to a failure to implement pharmacotherapy adequately in a number of trials (Hollon et al., 1991; Meterissian & Bradwejn, 1989). As a general rule, we think it is unwise to rely on meta-analyses unless something is known about the quality of the studies that have been included and there is confidence in the data.

A number of investigators have tried to deal with concerns of this kind by quantifying various aspects of design quality and incorporating these features in their analyses (Robinson et al., 1990; M. L. Smith & Glass, 1977). One aspect that consistently predicts variability across studies is the allegiance of the group conducting the investigation, a point we have previously mentioned. Although we applaud such efforts, we doubt that they are (as yet) sufficiently sophisticated to capture the complex interplay of multiple factors that need to be considered when evaluating the quality of the studies that make up a given literature. However, even if they were, the important point is that meta-analyses do not eliminate the need to make informed judgments about the quality of the studies reviewed, and, all too often, the people who conduct these analyses know more about the quantitative aspects of their task than about the substantive issues that need to be addressed.

Limitations of Efficacy

A final issue we ask evaluators to consider in describing a treatment's efficacy straddles the line between efficacy and effectiveness research. That is, for whom is the treatment beneficial? For example, has the treatment been shown to work only with well-educated clients or with clients from a single ethnic group? How does comorbidity affect efficacy? Practitioners often mention that research trials do not generalize well to their settings because research samples are vigorously screened to yield pure samples without comorbidity. From our reading of the literature, this is sometimes true but often not (cf. Wilson, in press).

A related issue concerns the interplay between clients' personal characteristics and treatment, that is, Aptitude × Treatment interactions (or moderator effects). In this exciting area of research, researchers attempt to identify variables such as personality characteristics that may affect reactions to a given treatment approach. Few consistent moderator effects have yet been established, perhaps as a result, in part, of the difficulty inherent in detecting interaction effects (B. Smith & Sechrest, 1991). Nonetheless, there are some intriguing early findings. For example, research by Beutler and colleagues (e.g., Beutler et al., 1991) and by Shoham and colleagues (e.g., Shoham, Bootzin, Rohrbaugh, & Urry, 1995) suggests that clients who are reactant (resistant) benefit more from nondirective therapy or paradoxical interventions than from standard cognitive or behavioral treatments. Such information is of considerable importance to practitioners, who must select among possible beneficial treatments for a particular client.

Effectiveness

In a recent report, the APA Task Force on Psychological Intervention Guidelines (1995) recommended that guidelines for treatment interventions be evaluated with respect to how closely they adhere to empirical evidence speaking to the effects of treatment. In their resultant template, the task force suggested that the greatest weight be given to the quality of the empirical evidence speaking to treatment efficacy, that is, whether the observed clinical change can be attributed to the treatment intervention. Movement along this efficacy dimension largely corresponded to the notion of internal validity, and RCTs or their logical equivalents were seen as the most powerful bases for drawing such inferences.

At the same time, the members of the task force recognized that there is more to determining whether a treatment has value than demonstrating that it can produce change under controlled conditions. They suggested that efforts to construct treatment guidelines should also take into consideration evidence of treatment utility or effectiveness, that is, whether the treatment can be shown to work in actual clinical practice. There is a growing recognition that controlled clinical trials may not capture the full richness and variability of actual clinical practice and a concern on the part of some that the very process of randomization may undermine the representativeness of the clinical encounter (e.g., Seligman, 1995). We do not necessarily share this latter concern, and we believe that randomization (or its logical equivalents) affords a particularly compelling means of testing for causal agency (i.e., whether observed change can be attributed to the treatment). Furthermore, like Jacobson and Christensen (1996), we believe that often RCTs could themselves be used to address many of their alleged deficits (e.g., Do patients randomly assigned to long-term treatment improve more than those assigned to time-limited treatment?). Nonetheless, we recognize that RCTs vary in the generalizability of their findings. At the effectiveness stage of research, we concur that quasi-experimental and nonexperimental designs can be fruitfully used to address questions of clinical utility (see also Hollon, 1996).[4]

Accordingly, we have asked evaluators to consider the effectiveness data for treatments they have determined to be possibly efficacious or established in efficacy and to include in their evaluation nonexperimental, quasi-experimental, and fully experimental research studies. Note, however, that we address here a particular and, as yet, very limited type of effectiveness research: the evaluation in clinical settings of a specified treatment already shown in controlled research to be efficacious.

[4] Although we frame our discussion in terms of *efficacy* and *effectiveness*, we think it unwise to draw too sharp a distinction between these terms. Rather, we prefer the more traditional distinction between internal and external validity and note that any given study (or body of literature) can be evaluated with respect to how informative it is on each dimension. Certain design features, particularly random assignment of participants to conditions, increase the confidence with which changes observed can be attributed to the treatment manipulations (internal validity). Nonetheless, there is no reason why controlled experimentation cannot take place in applied clinical settings using fully representative samples of patients and therapists, thereby also maximizing external validity. If the goal is to determine whether a particular treatment works in actual clinical practice (the question typically asked in effectiveness research), the most informative design is still an RCT conducted in that setting. Thus, we think the growing tendency to assume that all efficacy studies lack external validity or that effectiveness research must necessarily sacrifice internal controls is unfortunate. At the same time, we recognize that there are important questions regarding natural variations in patient characteristics and clinical practices that can best be addressed with purely observational methods and that, for some purposes, rigorous experimental control is neither necessary nor desirable. However, if the goal is to ascribe cause to an effect (i.e., to determine whether a treatment works in actual clinical practice), then the ideal design is one that maximizes both internal and external validity. Thus, the most informative designs for our current purposes are studies in which efficacy and effectiveness features converge.

Generalizability

Generalizability across populations. As defined by the task force, the clinical utility dimension not only incorporates the concept of external validity (generalizability) but also encompasses aspects of feasibility and cost utility. With respect to external validity, it is useful to consider the extent to which evidence from efficacy trials is relevant to the kinds of patients actually seen in clinical practice (cf. Krupnick, Shea, & Elkin, 1986). For example, there is a widespread belief that the patients studied in RCTs are necessarily less complex and easier to treat than the patients typically encountered in everyday clinical practice. However, this is not necessarily true. Although some RCTs have relied on analogue populations, others have sampled fully clinical populations. Participants in research trials do tend to be selected to be homogeneous with respect to the presence of a particular target problem and are often screened to be free of other more serious comorbid disorders, but inclusion criteria can be defined in an infinite number of ways. Most studies adhere to a rough severity hierarchy when defining their exclusionary diagnoses. For example, most studies of depression allow patients to be comorbid for personality or anxiety disorders but not schizophrenia, whereas most studies of schizophrenia also allow patients to be comorbid for depression. One need look no further than the work of Frank et al. (1990) with depressed patients with histories of multiple recurrences or the work of Linehan et al. with parasuicidal borderline patients (Linehan, Armstrong, Suarez, Allmon, & Heard, 1991) to find examples of controlled clinical trials with extremely difficult populations. There is nothing inherent in the logic of RCTs stating that the samples studied must be free from comorbid disorders or easy to treat. Many studies are done with fully clinical samples with whom it is extremely challenging to work.

Generalizability across therapists and settings. One factor that may boost outcomes in efficacy studies relative to clinical settings is therapist expertise. Therapists in RCTs often have access to a level of training and supervision not typically available to the average practitioner. Moreover, therapists in controlled trials often have the luxury of focusing on a specific type of problem that they can approach with a particular clinical intervention. We are intrigued with the development of specialized clinics for particular presenting problems in the public and private sector and expect that such clinics will be able to offer superior treatment because of their concentrated focus.

Another challenge made to the generalizability of findings from RCTs to clinical settings is the notion that the very act of controlling treatment changes its nature and threatens the practical utility of any findings obtained. Although we recognize the legitimacy of this concern (up to a point), we are intrigued that different commentators seem not to agree as to whether exercising greater control over what goes on in therapy is likely to make treatment more or less effective. For example, in their meta-analytic investigation of treatment benefits in clinical versus research settings, Weisz, Donenberg, Han, and Weiss (1995) reported better effects for treatment of children and adolescents in research settings than in clinical settings, and they tested a number of reasons why this might be the case (e.g., less disturbed patients in research settings and more frequent use of behavioral treatment methods in such settings). Persons (1991), on the other hand, has argued that the constraints imposed by controlled clinical trials prevent clinicians from using their judgment to tailor treatment to the idiosyncratic needs of their patients, implying that outcomes should be better in flexible treatment controlled by the practitioner. One of the few studies to examine this issue found that therapists produced better results when they followed a set protocol than when they tried to individualize treatment for each patient (Schulte, Kunzel, Pepping, & Schulte-Bahrenberg, 1992). Because the kind of therapy specified by protocol (exposure therapy) may be specifically efficacious for the phobic patients treated (most of whom met criteria for agoraphobia), it remains to be seen whether this finding will generalize to other types of patients and treatments. Nonetheless, this study does suggest that idiosyncratic clinical intuition may be no better guide to treatment modification than it is to clinical assessment (cf. Wilson, 1996).

All things being equal, those studies that most faithfully reproduce the conditions found in actual clinical practice are most likely to produce findings that generalize to those settings. It is not only important to know whether a treatment works in controlled research (efficacy); it is also important to know whether it works in a clinical setting under naturalistic conditions (effectiveness). However, differences between efficacy and effectiveness can result from any of a number of factors, some of which are fixed (e.g., the nature of the populations studied) and some of which are modifiable (e.g., the type of treatment used and the quality of the training and supervision provided). Naturalistic studies of actual clinical practice are certainly needed, but controlled trials (or their logical equivalents) are also needed to determine causal agency. The greater the clinical realism of those controlled trials, the greater their relevance to actual clinical practice.

Treatment Feasibility

Patient acceptance and compliance. In addition to issues of generalizability, the task force also recommended that issues related to feasibility be considered in evaluations of overall clinical utility. For example, many patients prefer systematic desensitization over exposure-based treatments even though the latter often work faster, largely because they also tend to generate more immediate distress (Masters, Burish, Hollon, & Rimm, 1987). Patients clearly have a right to choose the kind of treatment they receive, but clinicians have an obligation to ensure that their patients know the advantages and disadvantages of the various options available to them (see Pope & Vasquez, 1991). For example, most patients typically prefer psychological treatment over drugs in the treatment of depression, even though there is little empirical basis for choosing between them. If anything, the quality of the empirical support for the efficacy of pharmacotherapy is considerably greater than it is for the efficacy of psychological treatment (Muñoz, Hollon, McGrath, Rehm, & VandenBos, 1994). Likelihood of compliance is also an important issue; many patients fail to benefit from treatments that might otherwise be effective because they are unable or unwilling to adhere to the treatment regimen.

Ease of dissemination. Similarly, there are issues with respect to ease of dissemination. Even highly efficacious treatments are unlikely to be implemented if few people in clinical practice are competent to provide them. In this regard, it was disconcerting to learn that many clinical training programs and

internships largely ignore those treatments that have been empirically validated in favor of more traditional but less systematically evaluated approaches to treatment (Crits-Christoph, Frank, Chambless, Brody, & Karp, 1995). By way of contrast, when surveys of pharmacological practice revealed that many patients were either undermedicated (Keller et al., 1986) or being treated with the wrong drug (Wells, Katon, Rogers, & Camp, 1994), organized psychiatry responded with greater efforts at professional education (Regier et al., 1988) and the promulgation of practice guidelines (American Psychiatric Association, 1993a, 1993b; Depression Guideline Panel, 1993). Unfortunately, the response among the advocates of the more traditional psychotherapies has all too often been to dismiss the need for controlled clinical trials. We think that such a strategy is short sighted. Although we suspect that many of the largely untested strategies in current clinical practice would fare well in such trials, the time is rapidly approaching when unsystematic clinical impressions will no longer suffice to document a treatment's value, particularly when alternative treatments such as the pharmacotherapies exist that have been subjected to more rigorous empirical scrutiny.

At the same time, treatments that are straightforward and easier to learn are more likely to be disseminated to the larger practice community. For example, Jacobson and colleagues recently found that the simple behavioral activation component of cognitive therapy for depression was as effective as the full treatment package in which it is typically embedded (Jacobson et al., 1996). If this finding is valid, it could revive interest in more purely behavioral approaches to depression, because they are typically easier to master than the more complex cognitive interventions. All things being equal, one would expect that a simpler treatment intervention would be easier to disseminate than one that is more complex. Of course, all things are not necessarily equal, and existing proclivities and preferences may slow the penetration of even the simplest and most effective of treatments. Nonetheless, evaluators should be alert to those treatments that lend themselves to ease of dissemination either because they correspond to existing practices and predilections or because they are inherently less complex and easier to master.

Cost-Effectiveness

Finally, there is the issue of cost-effectiveness. Once again, all things being equal, those treatments that cost the least are likely to be preferred if there is no great difference in outcome. Nonetheless, decision making in this regard can be quite complex, particularly when the parties bearing the costs are not the ones receiving the benefits. Moreover, the situation becomes even more problematic when costs and benefits collide, that is, when the most effective treatment is also the most expensive.

Different treatments may differ in their cost-effectiveness as a function of time. For example, drugs typically cost less to provide than psychological treatment during the initial treatment period; they may prove more expensive over time, however, because they rarely have enduring effects that survive treatment termination (Hollon, 1996). Similarly, psychosocial interventions may also differ in this regard. Bulimic patients treated with interpersonal psychotherapy in a recent study showed less initial symptom change than patients treated with other more behavioral or cognitive interventions but continued to improve after treatment termination (Fairburn et al., 1993). More than a year later, patients previously treated with interpersonal psychotherapy were found to be doing as least as well as patients initially treated with cognitive–behavioral therapy (who tended to maintain their gains) and considerably better than patients initially treated with behavior therapy alone (who were particularly likely to relapse). This suggests that there may be an inverse relation between the exclusivity with which a treatment focuses on immediate symptoms resolution and the stability and breadth of the changes it obtains (Hollon, 1996). Regardless of the accuracy of this speculation, it clearly behooves evaluators of the literature to consider the relative costs and benefits of treatments not only in the short run but across the full life course of the patient, to the extent that the available information will allow.

Conclusions

We have touched on a great many issues of considerable complexity. On one hand, this will give readers a sense of the magnitude of the task that the evaluators undertook. On the other hand, we may have raised as many questions as we have answered, given the brevity of our coverage. We hope, in this case, that the references we have provided will be of assistance.

Necessarily, we have provided our view of important variables to consider in evaluating psychological treatment research, and others, including this special section's evaluators, might find much cause for disagreement. In particular, we recognize that not all would agree that randomized controlled clinical trials or their logical equivalents (e.g., single case experiments) represent the best (although not the only) means of detecting causal agency or that efficacy takes priority over effectiveness. Similarly, we expect that some will disagree with our decision not to require evidence of specificity to consider a treatment as efficacious. For this reason, although authors in this special section have been asked to start their reviews by examining the literature on the basis of the criteria we have delineated, they have also been invited to raise points of disagreement and to extend their reviews according to their own criteria. We hope that this will make for stimulating reading.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association. (1993a). Practice guidelines for eating disorders. *American Journal of Psychiatry, 150,* 792–794.

American Psychiatric Association. (1993b). Practice guidelines for major depressive disorder in adults. *American Journal of Psychiatry, 150*(Suppl. 4), 1–26.

American Psychological Association Task Force on Psychological Intervention Guidelines. (1995). *Template for developing guidelines: Interventions for mental disorders and psychological aspects of physical disorders.* Washington, DC: American Psychological Association.

Barber, J. P., Crits-Christoph, P., & Luborsky, L. (1996). Effects of therapist adherence and competence on patient outcome in brief dynamic therapy. *Journal of Consulting and Clinical Psychology, 64,* 619–622.

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). Elmsford, NY: Pergamon Press.

Beutler, L. E., Engle, D., Mohr, D., Daldrup, R. J., Bergan, J., Meredith,

K., & Merry, W. (1991). Predictors of differential response to cognitive, experiential, and self-directed psychotherapeutic procedures. *Journal of Consulting and Clinical Psychology, 59,* 333–340.

Beutler, L. E., Machado, P. P. P., & Neufeldt, S. A. (1994). Therapist variables. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 229–269). New York: Wiley.

Black, D. W., Wesner, R., Bowers, W., & Gabel, J. (1993). A comparison of fluvoxamine, cognitive therapy, and placebo in the treatment of panic disorder. *Archives of General Psychiatry, 50,* 44–50.

Burns, D. D., & Nolen-Hoeksema, S. (1992). Therapeutic empathy and recovery from depression in cognitive-behavioral therapy: A structural equation model. *Journal of Consulting and Clinical Psychology, 60,* 441–449.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., Pope, K. S., Crits-Christoph, P., Baker, M., Johnson, B., Woody, S. R., Sue, S., Beutler, L., Williams, D. A., & McCurry, S. (1996). An update on empirically validated therapies. *Clinical Psychologist, 49,* 5–18.

Christensen, A., & Jacobson, N. S. (1994). Who (or what) can do psychotherapy: The status and challenge of nonprofessional therapies. *Psychological Science, 5,* 8–14.

Clark, D. M., Salkovskis, P. M., Hackmann, A., Middleton, H., Anastasiades, P., & Gelder, M. (1994). Comparison of cognitive therapy, applied relaxation and imipramine in the treatment of panic disorder. *British Journal of Psychiatry, 164,* 759–769.

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., Luborsky, L., McLellan, A. T., Woody, G. E., Thompson, L., Gallagher, D., & Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1,* 81–91.

Crits-Christoph, P., Frank, E., Chambless, D. L., Brody, C., & Karp, J. (1995). Training in empirically validated treatments: What are clinical psychology students learning? *Professional Psychology: Research and Practice, 26,* 514–522.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology, 59,* 20–26.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity. *Psychological Bulletin, 52,* 281–302.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61,* 966–974.

Depression Guideline Panel. (1993). *Depression in primary care: Vol. 2. Treatment of major depression* (Clinical Practice Guideline No. 5, AHCPR Publication No. 93-0551). Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care and Policy Research.

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 57,* 414–419.

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

Fairburn, C. G., Jones, R., Peveler, R. C., Hope, R. A., & O'Connor, M. (1993). Psychotherapy and bulimia nervosa: Longer-term effects of interpersonal psychotherapy, behavior therapy and cognitive therapy. *Archives of General Psychiatry, 50,* 419–428.

Flick, S. N. (1988). Managing attrition in clinical research. *Clinical Psychology Review, 8,* 499–515.

Frank, E., Kupfer, D. J., Perel, J. M., Cornes, C., Jarrett, D. B., Mallinger, A. G., Thase, M. E., McEachran, A. B., & Grochocinski, V. J. (1990). Three-year outcomes for maintenance therapies in recurrent depression. *Archives of General Psychiatry, 47,* 1093–1099.

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 190–228). New York: Wiley.

Garfield, S. L. (1996). Some problems associated with "validated" forms of psychotherapy. *Clinical Psychology: Science and Practice, 3,* 218–229.

Greenhouse, J. B., Stangl, D., & Bromberg, J. (1989). An introduction to survival analysis: Statistical methods for analysis of clinical trial data. *Journal of Consulting and Clinical Psychology, 57,* 536–544.

Hollon, S. D. (1996). The efficacy and effectiveness of psychotherapy relative to medications. *American Psychologist, 51,* 1025–1030.

Hollon, S. D., & Beck, A. T. (1994). Cognitive and cognitive-behavioral therapies. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (pp. 428–466). New York: Wiley.

Hollon, S. D., & DeRubeis, R. J. (1981). Placebo-psychotherapy combinations: Inappropriate representations of psychotherapy in drug-psychotherapy comparative trials. *Psychological Bulletin, 90,* 467–477.

Hollon, S. D., Shelton, R. C., & Loosen, P. T. (1991). Cognitive therapy and pharmacotherapy for depression. *Journal of Consulting and Clinical Psychology, 59,* 88–99.

Jacobson, N. S., & Christensen, A. (1996). Studying the effectiveness of psychotherapy: How well can clinical trials do the job? *American Psychologist, 51,* 1030–1039.

Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., Gortner, E., & Prince, S. E. (1996). A component analysis of cognitive-behavioral treatment for depression. *Journal of Consulting and Clinical Psychology, 64,* 295–304.

Jacobson, N. S., & Hollon, S. D. (1996a). 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.

Jacobson, N. S., & Hollon, S. D. (1996b). Prospects for future comparisons between psychotropic drugs and psychotherapy: Lessons from the CBT vs. pharmacotherapy exchange. *Journal of Consulting and Clinical Psychology, 64,* 104–108.

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.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed.). Boston: Allyn & 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.

Keller, M., Lavori, P., Klerman, G., Andreasen, N. C., Endicott, J., Coryell, W., Fawcett, J., Rice, J., & Hirschfeld, R. (1986). Low levels and lack of predictors of somatotherapy and psychotherapy received by depressed patients. *Archives of General Psychiatry, 43,* 458–466.

Kendall, P. C., & Grove, W. M. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment, 10,* 147–158.

Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin, 65,* 110–136.

Klein, D. F. (1996). Preventing hung juries about therapy studies. *Journal of Consulting and Clinical Psychology, 64,* 81–87.

Krupnick, J., Shea, T., & Elkin, I. (1986). Generalizability of treatment studies utilizing solicited patients. *Journal of Consulting and Clinical Psychology, 54,* 68–78.

Lavori, P. W. (1992). Clinical trials in psychiatry: Should protocol deviation censor patient data? *Neuropsychopharmacology, 6,* 39–48.

Linehan, M. M., Armstrong, H. E., Suarez, A., Allmon, D., & Heard, H. L. (1991). Cognitive-behavioral treatment of chronically parasui-

cidal borderline patients. *Archives of General Psychiatry, 48,* 1060–1064.

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies: Is it true that "Everyone has won and all must have prizes"? *Archives of General Psychiatry, 32,* 995–1008.

Lyons, L. C., & Woods, P. J. (1991). The efficacy of rational-emotive therapy: A quantitative review of the outcome research. *Clinical Psychology Review, 11,* 357–369.

Masters, J. C., Burish, T. G., Hollon, S. D., & Rimm, D. C. (1987). *Behavior therapy: Techniques and empirical findings* (3rd ed.). New York: Harcourt Brace Jovanovich.

Meterissian, G. B., & Bradwejn, J. (1989). Comparative studies on the efficacy of psychotherapy, pharmacotherapy, and their combination in depression: Was adequate pharmacotherapy provided? *Journal of Clinical Psychopharmacology, 9,* 334–339.

Muñoz, R. F., Hollon, S. D., McGrath, E., Rehm, L. P., & VandenBos, G. R. (1994). On the AHCPR *Depression in Primary Care* guidelines: Further considerations for practitioners. *American Psychologist, 49,* 42–61.

Nicholson, R. A., & Berman, J. S. (1983). Is follow-up necessary in evaluating psychotherapy? *Psychological Bulletin, 93,* 261–278.

Orlinsky, D. E., Grawe, K., & Parks, B. H. (1994). Process and outcome in psychotherapy: noch einmal. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 270–376). New York: Wiley.

Parloff, M. B. (1986). Placebo controls in psychotherapy research: A sine qua non or a placebo for research problems? *Journal of Consulting and Clinical Psychology, 54,* 79–87.

Persons, J. B. (1991). Psychotherapy outcome studies do not accurately represent current models of psychotherapy: A proposed remedy. *American Psychologist, 46,* 99–106.

Pope, K. S., & Vasquez, M. (1991). *Ethics in psychotherapy and counseling: A practical guide for psychologists.* San Francisco: Jossey-Bass.

Regier, D. A., Hirschfeld, M. A., Goodwin, F. K., Burke, J. D., Lazar, J. B., & Judd, L. (1988). The NIMH Depression Awareness, Recognition and Treatment Program: Structure, aims and scientific basis. *American Journal of Psychiatry, 145,* 1351–1357.

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of the controlled outcome research. *Psychological Bulletin, 108,* 30–49.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565.

Schulte, D., Kunzel, R., Pepping, G., & Schulte-Bahrenberg, T. (1992). Tailor-made versus standardized therapy of phobic patients. *Advances in Behaviour Research and Therapy, 14,* 67–92.

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

Shoham, V., Bootzin, R. R., Rohrbaugh, M. J., & Urry, H. (1995). Paradoxical versus relaxation treatment for insomnia: The moderating role of reactance. *Sleep Research, 24,* 365.

Smith, B., & Sechrest, L. (1991). Treatment of Aptitude × Treatment interactions. *Journal of Consulting and Clinical Psychology, 59,* 233–244.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Strupp, H. H., & Hadley, S. W. (1979). Specific versus non-specific factors in psychotherapy. *Archives of General Psychiatry, 36,* 1125–1136.

Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically-validated psychological treatments: Report and recommendations. *Clinical Psychologist, 48,* 3–23.

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.

Weissman, M. M., & Markowitz, J. C. (1994). Interpersonal psychotherapy: Current status. *Archives of General Psychiatry, 51,* 599–606.

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63,* 688–701.

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117,* 450–468.

Wells, K. B., Katon, W., Rogers, B., & Camp, P. (1994). Use of minor tranquilizers and antidepressant medications by depressed outpatients: Results from the Medical Outcomes Study. *American Journal of Psychiatry, 151,* 694–700.

Wilson, G. T. (1996). Manual-based treatments: The clinical application of research findings. *Behaviour Research and Therapy, 34,* 295–315.

Wilson, G. T. (in press). The clinical utility of randomized controlled trials. *International Journal of Eating Disorders.*

Appendix

Summary of Criteria for Empirically Supported Psychological Therapies

1. Comparison with a no-treatment control group, alternative treatment group, or placebo (a) in a randomized control trial, controlled single case experiment, or equivalent time-samples design and (b) in which the EST is statistically significantly superior to no treatment, placebo, or alternative treatments or in which the EST is equivalent to a treatment already established in efficacy, and power is sufficient to detect moderate differences.

2. These studies must have been conducted with (a) a treatment manual or its logical equivalent; (b) a population, treated for specified problems, for whom inclusion criteria have been delineated in a reliable, valid manner; (c) reliable and valid outcome assessment measures, at minimum tapping the problems targeted for change; and (d) appropriate data analysis.

3. For a designation of efficacious, the superiority of the EST must have been shown in at least two independent research settings (sample size of 3 or more at each site in the case of single case experiments). If there is conflicting evidence, the preponderance of the well-controlled data must support the EST's efficacy.

4. For a designation of possibly efficacious, one study (sample size of 3 or more in the case of single case experiments) suffices in the absence of conflicting evidence.

5. For a designation of efficacious and specific, the EST must have been shown to be statistically significantly superior to pill or psychological placebo or to an alternative bona fide treatment in at least two independent research settings. If there is conflicting evidence, the preponderance of the well-controlled data must support the EST's efficacy and specificity.

Received October 14, 1996
Revision received March 1, 1997
Accepted March 17, 1997 ■