Research on Psychotherapy Efficacy and Effectiveness: Between Scylla and Charybdis?

Peter E. Nathan, Scott P. Stuart, and Sara L. Dolan
University of Iowa

Clinical researchers have recently begun to explore differences between psychotherapy outcome studies that focus on efficacy and those that focus on effectiveness. The authors provide concise descriptions of these research models, followed by more extended consideration of the most important conceptual and empirical distinctions between the two. Research on the efficacy/effectiveness distinction is then put into context: The common treatment variables that also influence treatment outcomes are reviewed. Fifty years of research on psychotherapy outcomes are next considered; contemporary research on the efficacy and effectiveness research models is emphasized. A description and evaluation of current efforts to heighten the value of technique-focused research to clinicians follow. The authors conclude by anticipating some promising future directions in this research domain.

Psychotherapy 50 years ago and psychotherapy today share very little. Available then only to a privileged few patients, psychotherapy was offered by the few psychiatrists and fewer psychologists and others trained at the minute number of available training centers. Fifty years ago, psychotherapy came in only one variety, heavily flavored by Freudian theory and practice. As a result, psychotherapy patients had to make a substantial investment in both time and money, so they were most often members of the elite financially and psychologically.

Psychotherapy training today is offered by virtually every clinical and counseling psychology training program, most psychiatric residencies, and many training programs in clinical social work and psychiatric nursing. It is the object, moreover, of many other pre- and postdoctoral training programs designed for core mental health professionals or for other helping professionals. Although managed care has succeeded in recent years in reducing access to long-term psychotherapy, it remains widely available in some form to most strata of society. Because the drug revolution has led many psychiatric residency programs to de-emphasize training and research in psychotherapy in favor of psychopharmacology, it is largely in clinical and counseling psychology training programs and laboratories that today’s most significant efforts to explore and develop psychotherapy further are taking place. Much of this research focuses on behavior therapy and cognitive–behavioral therapy (CBT), which have succeeded psychodynamic therapy as the psychotherapeutic treatments of choice for many conditions, primarily because of the strong empirical support they have garnered for their efficacy.

This review of 50 years of psychotherapy research ultimately highlights a distinction in psychotherapy research tactics that has attracted increasing attention in recent years. The distinction is between research on efficacy (e.g., outcome assessment under conditions of high internal validity) and research on effectiveness (e.g., outcome assessment under conditions of high external validity). Some researchers claim that only outcome research that takes place under laboratory conditions and follows the efficacy model provides meaningful psychotherapy outcome data, others swear by the real-world effectiveness model, and still others believe that only integration of the two will yield valid data on efficacy. This comprehensive historical review of the literature on this issue seeks to determine whether this question can be answered by research to date on the differential utility of psychotherapy efficacy and effectiveness research.

Psychotherapy Efficacy and Effectiveness Research

Efficacy research focuses on the measurable effects of specific interventions. The clinical trial represents the prototypic efficacy study: it compares one or more experimental treatments with one or more control or comparison treatments that can be standard treatment, placebo, or waiting list. To maximize the likelihood of detecting treatment effects, factors that might obscure them are eliminated in efficacy studies to the extent possible. To reduce the impact of outcome expectations that participants might hold, clinical trials of psychotherapy are often single-blinded, meaning that patients do not know whether they have been assigned to an experimental, comparison, or control condition. (Double-blinding, in which the therapist or experimenter is also unaware of the identity of the active treatment, is generally a feature only of outcome studies of pharmacological treatment because of the near impossibility of disguising the intent of a psychotherapeutic intervention.) Efficacy studies also tend to use restrictive inclusion and exclusion criteria to create the homogeneous groupings of participants that make it possible to link a specific efficacious treatment to a particular diagnostic entity. Because not all potentially confounding factors can be identified or controlled, participants in efficacy studies are commonly randomly assigned to experimental and control or comparison conditions to reduce the likelihood that those conditions will differ in unanticipated ways. Efficacy trials
also typically use clinicians who are rigorously trained in the treatment they are to provide (it might also be incorporated in a treatment manual) to ensure that the treatment provided is the treatment intended. Treatment in efficacy studies is often provided at no cost to patients. Outcomes of treatment in efficacy studies are most often assessed on a short-term, targeted basis, typically focused on changes in symptoms rather than on more global changes in personality or quality of life.

Effectiveness research aims to determine whether treatments are feasible and have measurable beneficial effects across broad populations and in real-world settings. Typically included in effectiveness studies are persons in need of treatment, regardless of specific diagnoses, comorbid psychopathology, or length of illness. Clinicians in these studies are generally not trained specifically in the research protocol. Moreover, clinical circumstances rather than research design most often dictate choice of treatment method, frequency, duration, and outcome assessment. Although effectiveness studies may randomize assignment of patients to treatments, disguising the treatment to which a patient has been assigned is often not feasible. Outcome measures are often broadly defined as changes in degree of disability, quality of life, or personality, for instance, rather than targeted evaluations of clinical status.

Barlow (1996) has provided succinct definitions of efficacy and effectiveness research. Efficacy refers to "the results of a systematic evaluation of the intervention in a controlled clinical research context. Considerations relevant to the internal validity of these conclusions are usually highlighted" (p. 1051). Effectiveness has to do with "the applicability and feasibility of the intervention in the local setting where the treatment is delivered" and is designed to "determine the generalizability of an intervention with established efficacy" (p. 1055). Efficacy studies emphasize internal validity and replicability; effectiveness studies emphasize external validity and generalizability.

Factors That Distinguish Efficacy and Effectiveness Studies

A range of differences in research methodology differentiates efficacy and effectiveness studies. For the most part, as the historical review of psychotherapy research that follows indicates, the specific factors that differentiate efficacy and effectiveness studies have been studied intensively only during the past decade, reflecting the maturation in research methodology that has taken place during that time.

Efficacy studies are concerned above all with replication. They are designed so that other researchers can conduct similar studies in similar settings to test the same hypotheses. This focus leads naturally to a number of research design issues that are of much less importance to those conducting effectiveness studies. Primary among these is the need to construct an appropriate control condition with which the experimental condition can be compared. As a corollary, the emphasis on internal validity dictates the need for random assignment of subjects to either an active or a control condition. Efficacy studies must also define the type of treatment being provided so that the therapy can be replicated as closely as possible in subsequent studies. This consideration has led to the widespread use of treatment manuals in efficacy studies, as well as procedures to ensure that the therapists conducting the intervention adhere reliably to the dictates of the manual. As well, priority is given in efficacy studies to well-defined groups of patients whose problems are specified using objective measures of pathology.

Appropriate Control Conditions

The inclusion of a control condition in psychotherapy studies raises two fundamental questions. The first involves the theoretical issue of whether or not a control condition is needed to draw valid conclusions about whether a treatment is beneficial. The second is the nature of the control condition when one is used. Theorists have long debated whether a placebo control is even possible in psychotherapy studies, given that the nonspecific factors present in almost any credible psychotherapy placebo have been identified by many experts to be powerful agents of change (J. D. Frank, 1971; Greenberg, 1994; Orlinsky & Howard, 1975).

There is little debate on the need for control conditions in outcome studies designed to maximize internal validity. In their discussion of efficacy and effectiveness studies, Mintz, Drake, and Crits-Cristoph (1996) concluded that both should include well-specified control conditions. Several theoretical objections to the inclusion of control conditions have been raised, however. These include an ethical objection to subjecting clients to a treatment presumed to be inert (or relatively less effective than the treatment being investigated) and a concern about the biases in treatment that may be introduced by the inclusion of a control group. The ethical objections have been addressed most recently with the general acknowledgment that placebo or presumed inert control conditions are not needed if the experimental treatment can be compared with a treatment of known effectiveness. In part because of these concerns, both Parloff (1986) and P. Horvath (1988), among others, have strongly suggested that psychotherapy outcome studies compare a treatment of known benefit with the experimental treatment, reserving placebo controls for studies in which purported mechanisms and processes of change are under investigation.

The second objection to the inclusion of control conditions is that they introduce bias in the patient population. For example, Seligman (1995) has criticized use of randomized control conditions. His point is that patients in community settings enter treatment in an active fashion to seek treatment for self-perceived problems, as opposed to research patients who enter treatment passively and must accept the possibility that they may be assigned to an ineffective treatment. Though internally consistent across studies, the passive entrance of research patients likely biases efficacy studies toward patients who are fundamentally different from those who seek treatment in effectiveness studies in naturalistic clinical settings.

Psychotherapy placebos can be divided into three broad categories, with the caveats that they may be used differently across studies and that the conceptualizations of each may vary across paradigms as well. Placebos may be characterized as completely inert, theoretically inert, or intentionally including nonspecific factors (discussed in the next section of this article) that may influence outcome but that are missing the therapeutic elements hypothesized to affect outcome in the experimental treatment.

Those conceptualized as inert are typically designed to provide a treatment that has no therapeutic benefit but retains a reasonable degree of plausibility. To serve as a placebo, the control condition
must be plausible to both the patient and the clinician, and thus, it must contain some elements of optimism for improvement, offering patients at least the possibility that they might improve. Psychotherapy placebos must meet more stringent requirements in this regard than medication placebos, as the therapist cannot be blinded to treatment. To create conditions in which the therapist is invested in the treatment, a rigorous rationale must be provided for the intervention. In this case, psychotherapy placebos presumed to be inert contain, in fact, many of the nonspecific factors believed to influence outcome, such as the provision of hope (J. D. Frank, 1973) and positive regard from the therapist providing the treatment (Rogers, 1957). Thus, despite the inert status accorded to some placebo controls, most are in fact active, if not specific, treatments. Those that are truly inactive typically lack plausibility and are therefore given questionable status as placebos.

Studies that use theoretically inert placebos include control interventions thought to be ineffective treatments for the disorder being studied. The National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH-TDCRP; Elkin et al., 1989), detailed later in this article, is a classic example of such a study. Besides constituting a control for the pharmacological effects of medication, the placebo medication condition was designed to lack any of the hypothesized mechanisms of action of either interpersonal psychotherapy (IPT; Klerman, Weissman, Rounsaville, & Chevron, 1984) or CBT (Beck, Rush, Shaw, & Emery, 1979). Great care was taken to ensure that the clinicians who provided the placebo treatment did not use techniques of either IPT or CBT and that they avoided such interventions reliably (Elkin, Parloff, Hadley, & Antry, 1985). Nonetheless, patients in the placebo medication condition reported symptom improvement comparable to that experienced by patients in the active treatment groups.

The NIMH-TDCRP exemplifies a problem inherent in this particular type of placebo design, in that some critics of the study have interpreted its findings to reflect, in substantial part, the impact of nonspecific effects. Many authors believe that nonspecific effects, irrespective of the presumed mechanism of action of an active treatment, are responsible for a great deal of the outcome variance in psychotherapy studies. Such factors as the provision of hope for change (J. D. Frank, 1971), the effects of persuasion (J. D. Frank, 1973), the opportunity to interact with a caring clinician (Rogers, 1957), and the demand characteristics of therapy (P. Horvath, 1988) are all potential agents for change in any credible placebo intervention. It has also recently been suggested that elements intrinsic to efficacy studies, such as the increased amount of contact with research personnel that typically occurs (Thase, 1999a) and the influence of the questionnaires used in the study on its outcomes, may also effect change.

Concerns over control of these influences have led researchers to adopt a third type of control condition. In it, the control or comparison treatment is acknowledged to contain nonspecific elements and is specifically constructed to emulate the active treatment in question, with the exception that it is designed not to include the active elements being tested in the active treatment. This type of deconstructed or dismantled comparison treatment has been used most frequently in psychotherapy research. A typical paradigm might involve the test of a behavior therapy intervention for panic disorder such as relaxation. In this situation, the comparison treatment would mimic all of the elements of the experimental treatment, including amount and frequency of contact, and would include even nonrelaxation elements such as exposure and desensitization; it would not include the relaxation induction. Although such a control condition might resolve questions about mechanisms of change, it is not likely to shed much light on the efficacy of the experimental treatment.

Two other types of nonplacebo control conditions have also been used in both efficacy and effectiveness studies. In some research, comparison of an active treatment with a treatment-as-usual condition is conducted. This type of control condition may be very appropriate for determining whether the treatment being evaluated is better than what is currently provided in the community—a question often asked by effectiveness researchers. Such a comparison may not, however, meet the criteria noted above for demonstrating efficacy of a treatment—namely, that the experimental treatment is equal to or better than a known efficacious treatment, as usual treatment may not be of proven efficacy and thus an invalid comparator. The treatment-as-usual paradigm, however, is frequently used as a placebo treatment in efficacy studies rather than as an active effectiveness comparator. In this situation, it would appear to be a very poor placebo, as it contains all of the nonspecific elements found in psychotherapeutic treatments, may include elements of the presumed active treatment, and is of questionable internal consistency.

Some authors have also used waiting list controls for psychotherapy studies. In these types of designs, no attempt is made to provide patients with a plausible placebo treatment. In fact, patients are advised that the waiting list is likely to be ineffective. Though elements such as increased contact with other research personnel may be present, researchers conceptualize the waiting list as nonspecific. The drawback to this approach is that it may produce the type of patient bias noted above by Seligman (1995): Patients who enroll in psychotherapy studies in which they may be randomized to a waiting list control condition may be very different from those who seek treatment in community clinics.

Random Assignment of Participants

Random assignment to treatment is typically done whether the experimental treatment is being compared with a placebo or with another active treatment. Prospective, randomized assignment is considered an essential element of methodologically rigorous efficacy research (Mintz et al., 1996) because it provides statistically unbiased estimates of treatment effects. Effectiveness studies, on the other hand, may utilize post hoc comparisons of disaggregated groups, such as a retrospective comparison of depressed patients who received psychotherapy with those who received medication.

Random assignment to treatment has been heavily criticized for introducing a selection bias, in that clients may self-select for these types of efficacy studies (Seligman, 1995). However, though it is theoretically appealing to assume that patients who seek treatment actively in community settings are different from those who enter research protocols and face the prospect of randomization, there are no data to date that have substantiated this view. Hence, a more compelling critique is the argument that community-based effectiveness research may not allow for prospective randomization because particular clinic sites may be required to provide an active intervention, chosen by the patient in consultation with his or her clinician, to all who seek treatment. Moreover, if the goal of the
effectiveness research is service oriented, that is, designed to quantify the effects of a given treatment for those patients who actively seek it, use of random assignment is obviously not desirable.

Random assignment to treatment does, however, appear to have an impact on outcome in research studies. In a meta-analysis of studies of marital and family psychotherapy, Shadish and Ragsdale (1996) reported that studies using random assignments to treatment had consistently higher mean posttest effects and less variable posttest effects than studies using nonrandom assignment. Though the authors found that the differences between the two groups of studies could be reduced somewhat by considering various covariables, the differences persisted, leading the authors to conclude that random assignment should continue to be considered the gold standard in psychotherapy efficacy studies.

Treatment Manuals

Psychotherapy manuals have become an essential part of most efficacy studies, in large part because it is generally agreed that efficacy studies should include at least the following three components: (a) a definitive description of the techniques specific to the therapy being tested, (b) a clear statement in concrete terms of the operations the therapist is to perform, and (c) a measure of the degree to which the therapist adheres to the prescribed techniques (Kiesler, 1994). Many therapy manuals describe the therapeutic stance to be taken by the clinician and also specify what types of techniques are not permissible. These measures facilitate a more objective comparison of differing psychotherapies and improved training of therapists in the manualized therapy (Kiesler, 1994).

Eifert, Schulte, Zvolensky, Lejuez, and Lau (1997) cited a number of advantages of manualized research paradigms, prominently including that they increase internal validity and standardization. In addition, Eifert et al. noted that manuals are helpful for clinical practice because they enable clinicians to administer efficacious treatments more effectively, reduce idiosyncrasy in therapeutic methods, and increase the clinician’s focus on specific treatment goals and techniques. Detractors of manualized therapies have focused on theoretical weaknesses of the manual approach and on concerns that manuals prevent clinicians from fully practicing their therapeutic skills by constraining their possible responses and making it more difficult for them to use client feedback. To this end, when Crits-Christoph (1996) reexamined the results of his earlier meta-analysis of therapist effects in psychotherapy outcome studies (Crits-Christoph et al., 1991), he reported that manuals did “reduce therapist differences to a rather low level” (p. 260), specifically, to 4% of outcome variance due to therapist differences, as against 13% in studies that did not use treatment manuals.

The principal theoretical objection to therapy manuals is that they are nearly always disorder based (Eifert et al., 1997) rather than theory driven, because they typically use Diagnostic and Statistical Manual (DSM) classification to determine criteria for the disorder being treated. Both CBT and IPT are excellent examples of this trend, with a proliferation of modified therapies being developed for a wide variety of DSM–IV (American Psychiatric Association, 1994a) disorders once efficacy for the treatment of major depression was established.

Another theoretical argument against manual-based research is that it fails to inform clinicians about how to effectively treat specific patients. Two variations of the argument have been proposed. The first is that empirical findings are based on group means, and the average patient for whom such findings apply is simply a statistical abstraction that does not conform to the individual patients actually treated (Macln & Howard, 1994). The second avers that having to adhere to treatment manuals prevents therapists and patients from working together in a self-correcting way to improve outcome (Seligman, 1995). It is true that typical efficacy studies require that therapists follow a strictly defined protocol, with no little or no allowance for the exercise of clinical judgment. Although most studies allow some flexibility in what can be done within a given session, therapists are not allowed to change the scheduling or frequency of sessions prescribed by the protocol (except to schedule well-defined emergency sessions when indicated).

A final objection to therapy manuals is the art versus science argument. It is well articulated by Edelson (1994), who argued that psychotherapy is an art based on relativistic observations. Rather than emphasizing reliability, the clinician strives to make meaningful observations uniquely based on the therapeutic dyad. The practice of psychotherapy, in his view, is concerned with the particulars of a given client as an individual, whereas psychotherapy research is concerned with generalities drawn from a population.

There is some evidence that the use of treatment manuals in efficacy studies does lead to differences in outcome. In a meta-analysis of short-term psychotherapies, Anderson and Lambert (1995) found that studies using treatment manuals had larger effect sizes than those that did not. This conclusion, however, may have been biased. The efficacy studies in question emphasized adherence to a treatment manual; studies allowing therapists to choose either to use their clinical judgment in a naturalistic treatment setting or to adhere to a treatment manual might have provided more objective data on outcomes.

Most efficacy studies use some measure of therapist adherence to the treatment manual as a means of preventing therapist drift. There is some evidence that adherence is correlated with greater therapy effects. E. Frank, Kupfer, Wagner, McEachran, and Cornes (1992) found that therapist adherence to IPT was directly correlated with improved outcome. However, when Jacobson and his colleagues (1989) compared the results of two versions of behavioral marital therapy delivered by therapists whose therapeutic flexibility was either constrained or encouraged, outcomes did not differ, although a trend for less relapse at follow-up in the flexibly treated couples was reported.

As a practical consequence of the need to maintain therapy integrity, manual-driven therapies tend to be short term, given that therapeutic drift becomes more problematic with the passage of time. As a consequence, and with some notable exceptions (e.g., E. Frank, Kupfer, & Perel, 1990; Koenberg, Selzer, Koenigsberg, Carr, & Appelbaum, 1989; Linehan, 1993), most efficacy studies are designed to test time-limited treatments for acute disorders. Coincidentally, the time-limited nature of efficacy studies parallels changes in clinical practice under managed care. Today’s limited insurance benefits frequently constrain number of therapy sessions and may require therapist and client to negotiate ways to stretch a limited number of sessions over a lengthy period of time.
Well-Defined Groups of Patients

Efficacy studies require the reliable evaluation of patients and the use of reliable measures of symptoms over time (Minz et al., 1996). Typically, potential patients are screened using standardized diagnostic instruments, generating disorder-based groups to whom the intervention can be provided. Replicability and internal validity requirements also dictate that patients selected for treatment be as homogeneous as possible in their diagnoses. The goal, which is not always realized because comorbid conditions are not always identified, is the generation of treatment and control groups with few or no confounding diagnoses. As many as 5 to 10 potential participants may be screened for every one enrolled in more rigorous studies (Thase, 1999a), both because of potentially confounding comorbidity and because their disorders may not be sufficiently severe.

Great care is also taken in efficacy studies to ensure that measures of symptomatology are valid and reliable. The ideal study combines measures from three observation points. These include patient self-reports, ratings from blinded and objective observers, and psychophysiological measures, if available. Although ratings from treating therapists may also be used, they have a distinct bias potential. All assessments are cast in quantifiable terms so that comparisons can be made. Several different measures of a given pathological construct, such as depression, are also used. The measurement of longitudinal change requires multiple assessments over time, with careful consideration of the times at which the measures are obtained (Gottman & Rushe, 1993).

Critics have noted several concerns about these assessment provisions. First, measures of psychopathology are usually chosen because of their conformation to hypothesized mechanisms of change or active ingredients (Elkin, Pilkonis, Docherty, & Sotsky, 1988); some measures may also be chosen to reflect a theoretical conceptualization of the disorder being studied. In either case, the information gleaned from the treatment trial is limited by the assessment instruments used; instruments that are more restricted conceptually may miss changes in other domains because they do not measure them. Each school of psychotherapy may have its own unique instruments, making comparisons across studies very difficult.

Moreover, the measures used in a study may create bias in the response of patients to the interventions under investigation. In most cases, this demand bias is equivalent across the interventions being tested in a given study. However, there are some cases in which a treatment being assessed may be compromised by assessment questionnaires. For example, the effectiveness of a psychological treatment designed to treat somatization disorder and specifically constructed to avoid discussion of somatic symptoms by redirecting the patient away from discussion of physical problems may be impaired when patients are asked to respond in detail to repeated questions about physical symptoms. Others have questioned the limitation imposed by objective measures of outcome on the range of possible outcomes. To this end, attempts have been made to expand the constructs that are measured, including such elusive concepts as quality of life and life satisfaction. Many clinicians, however, continue to be dissatisfied with these efforts. They argue that research measures assess illness rather than suffering, do not capture the subjective reality of the patient, and do not address issues important to more psychodynamically inclined therapists, such as ego functioning and defense mechanisms (Edelson, 1994; Greenberg, 1994; Spence, 1994).

Research on Common Factors

Are differences in psychotherapy outcomes strongly associated with specific types or schools of psychotherapy, as the historical overview of psychotherapy research that follows suggests? Or do variables common to all psychological treatments have the greatest influence on outcomes, as the infrequent report of no differences in outcomes among treatments (called by some the “dodo bird effect”) implies? In fact, through the years, a number of well-respected investigators have concluded that the so-called common variables may carry a very substantial amount of the treatment outcome variance. That research is reviewed briefly here.

Lambert and Bergin (1994) located common factors in the therapist, the client, and the therapeutic process. Within each of the three, they added, are three subfactors: support factors, such as therapeutic alliance, catharsis, and therapist warmth; learning factors, including corrective emotional experience, insight, and feedback; and action factors, which include cognitive mastery, modeling, and behavioral regulation. Many of these variables, these authors claimed, affect outcomes independent of specific therapeutic techniques employed.

According to Beutler, Machado, and Neufeldt (1994), the therapist variables examined for their impact on outcomes extend across a broad continuum from objective demographic characteristics and sociocultural background factors of psychotherapists to subjective factors like therapist values, attitudes, and beliefs. They include factors quite specific to therapy (e.g., the therapist’s role in the therapeutic relationship and his or her expectations for its success) as well as those well removed from it, such as cross-situational variables (e.g., therapists’ cultural attitudes and emotional well-being). Therapist variables reflecting therapy-specific states, including the therapist’s professional background (see, e.g., Berman & Norton, 1985) and his or her style and choice of interventions (Robinson, Berman, & Neimeyer, 1990), appear to exert the most powerful effects on therapy outcomes. Therapist-specific variables such as age and gender (see, e.g., Zlotnick, Elkin, & Shea, 1998) seem to affect therapy outcomes the least.

With occasional exceptions (see, e.g., Luborsky & Diguer, 1995), patient variables have not shown a consistent, robust relationship to therapeutic outcomes. Although the NIMH-CDCRP (Elkin et al., 1989) carefully screened potential participants for specific defined characteristics both to maximize diagnostic homogeneity and to facilitate treatment effectiveness, no single patient variable correlated robustly with outcome. In the same vein, although the YAVIS variables (patient youth, attractiveness, verbal ability, intelligence, and success in other domains) were once thought to be robust predictors of psychotherapy treatment success (see, e.g., Stoler, 1963), they have since been largely discounted. Other kinds of patient variables have attracted more recent empirical interest, including initial level of disorder severity and degree of psychiatric comorbidity (Elkin, 1995). It is largely from patient variables like these that empirical support is now sought for the hypothesis that matching patients to specific treatments based on patient attributes may produce more robust treatment effects (see, e.g., Project MATCH Research Group, 1993, 1997).
In 1973, Hans Strupp (1973) reported that therapeutic process variables—factors influencing therapists’ reactions to patients’ behavior and attitudes and vice versa—directly affect therapeutic outcomes. In a 1986 review of process-outcome research, Orlinsky and Howard (1986) concluded that process variables, including the strength of the therapeutic bond, the skillfulness with which interventions are performed, and the duration of the treatment relationship, all had a positive impact on outcomes. Nonetheless, critics of process research have continued to emphasize the difficulties of studying relationships between therapeutic process and outcome. Many (e.g., Butler & Strupp, 1986; Elkin, 1995; Stiles & Shapiro, 1989) have lamented the recent trend to design therapy outcome studies like drug trials (the so-called drug trial metaphor), given the problem in trying to isolate the active (process) ingredients of psychotherapy.

Following an extensive review of outcome studies, Lambert (1992) suggested that about 30% of psychotherapy outcome variance is attributable to therapist variables affecting the relationship with the patient, such as empathy, warmth, and acceptance of the patient. Svarterg, Selzter, and Stiles (1998), Horvath and Symonds (1991), and Eaton, Abeles, and Gutfriend (1988) have concluded that the therapeutic alliance is the most important factor in determining positive therapeutic outcomes. Much earlier, Carl Rogers (1961) came to the same conclusion when he identified the necessary and sufficient conditions for a productive helping relationship as being therapist’s unconditional positive regard for the patient, empathy, and genuineness. Ten years later, Hans Strupp (1973) also observed that the therapist’s attitudes and feelings toward the patient have the potential powerfully to affect therapeutic outcome.

A recent resurgence in research on therapists’ comfort with and belief in the therapy they practice has taken place. In comparative studies of psychotherapies, for example, the therapist’s allegiance to a particular brand of psychotherapy has turned out to be a potent predictor of positive outcome. Luborsky et al. (1999) reviewed 29 treatment comparison studies, assessing researchers’ allegiance to a compared treatment three ways: assessment of allegiance to a specific treatment as expressed in published studies (reprint-based assessment), Likert-type ratings of allegiance by close colleagues, and Likert-type self-ratings. The three methods were moderately intercorrelated, and all three positively related to treatment success. Overall, a composite measure of allegiance was found to correlate significantly with outcome, explaining 69% of its variance. Some observers, however, believe that this effect is inflated by researchers’ commitment to a specific treatment modality (Holton, 1999; Thase, 1999b). Others urge caution in interpreting these results lest the original goal of comparative studies be lost sight of: to find out what treatments work, for what people, under what conditions (Klein, 1999; Lambert, 1999; Shoham & Rohrbaugh, 1999).

Overview of Psychotherapy Outcome Research

From the 1950s to the 1980s

The 1950s and 1960s: Inadequate Research Methods, Disappointing Efficacy, Significant but Meaningless Psychotherapy Effects

Although Hans Eysenck’s 1952 evaluation of the effects of psychotherapy, elaborated in a 1960 review (Eysenck, 1952, 1960), was not the first to assess psychotherapy outcomes, its emphasis on the primacy of data and its willingness to reach an unpopular conclusion distinguished it from other early evaluations of psychotherapy. The review’s two most notable conclusions: (a) Approximately two thirds of so-called neurotic patients recover or improve markedly within 2 years of the onset of their disorder, in the absence of treatment; and (b) there appears to be an inverse correlation between recovery and psychotherapy. Eysenck’s controversial 1952 bottom line, elaborated in 1960: The psychotherapies in widest use at midcentury were largely ineffectual.

Of specific relevance to the focus of this review was Eysenck’s acknowledgment in both the 1952 and 1960 reviews, in response to numerous critics, that the inadequate methodology of the outcome studies on which he based his conclusions required qualification of his most provocative conclusions. Few of the studies included in Eysenck’s (1952) review were controlled; the nature and severity of patients’ illnesses were inadequately described, in part because of the unreliability of DSM-I (American Psychiatric Association, 1952); the length and precise nature of the treatments provided were insufficiently detailed; and treatment follow-ups were generally inadequate. In other words, virtually all the research on psychotherapy that Eysenck reviewed in 1952 failed to meet the methodological criteria today considered minimally necessary before including a study in a review of efficacy or effectiveness research. Accordingly, the methodological limitations of the studies on which Eysenck based his 1952 conclusions render its findings of little or no utility by current standards.

Eysenck’s (1960) reexamination of the psychotherapy outcome literature 8 years later was based on data from research that was somewhat more adequate methodologically. He examined psychotherapy outcomes from the Cambridge—Somerville delinquency prevention project (Powers & Witmer, 1951; Teuber & Powers, 1953), for its time a state-of-the-art prevention/treatment study, and Rogers and Dymond’s (1954) investigation of nondirective therapy, which broke new ground by using a control group and multiple, newly developed outcome measures. Nonetheless, Eysenck’s conclusions in 1960 differed little from those in 1952: “With the single exception of psychotherapeutic methods based on learning theory, results of published research with military and civilian neurotics, and with both adults and children, suggest that the therapeutic effects of psychotherapy are small or non-existent” (Eysenck, 1960, p. 245).

Initiating our historical review of the psychotherapy outcome literature with Eysenck’s 1952 and 1960 evaluations is intended to emphasize both the primitive state of psychotherapy research at midcentury and its discouraging findings on therapeutic outcomes. The design shortcomings ensured that the methodological subtleties on which current distinctions between efficacy and effectiveness research are based, as outlined above, were simply unavailable to the researchers of the time. Nonetheless, the broader problem that the efficacy/effectiveness distinction epitomizes—the real-world irrelevance of much psychological research, including research on psychotherapy outcomes—was very much on the minds of psychologists very early.

In his well-known presidential address to the American Psychological Association describing the “two disciplines of scientific psychology,” for example, Cronbach (1957) distinguished between correlational methodologies (often used in clinical research) and experimental methods (more easily employed in laboratory set-
tngs), concluding that only the latter detect causal agency. A few years later, writing in an influential statistics text, Hays (1963) affirmed that

all significant results do not imply the same degree of true association between independent and dependent variables. Virtually any study can be made to show significant results if one uses enough subjects, regardless of how nonsensical the content may be. (p. 326)

Two years afterward, elaborating on Hays’s observation from the clinical perspective, Cohen (1965) made one of the earliest references to an often-repeated truism about significant yet meaningless psychotherapy outcome research: Findings that are statistically significant may nonetheless be insignificant clinically.

Taking a distinctly contrary position for this era, Bergin (1966) identified six notable findings from psychotherapy research that he felt had clear relevance to clinical practice. They included the following: (a) Psychotherapy changes behavior, attitudes, and adjustment; (b) controls may improve over time as a result of “informal therapeutic encounters” (p. 225); (c) progress in therapy is a function of therapeutic warmth, empathy, adjustment, and experience; (d) only client-centered therapy has been validated empirically; (e) the psychodynamic therapies are of limited effectiveness and affect only a small number of conditions; and (f) behavior therapies show considerable promise.

The 1970s: Improved Research Methods, Enhanced Efficacy, Not Much More Help for Practitioners

A series of comprehensive reviews of the psychotherapy research literature in the early 1970s reflected two themes: hard-won advances in psychotherapy research methodology and encouraging data on the efficacy of new therapies, especially the behavior therapies and CBTs. Surprisingly, however, a number of these reviews, including Bergin and Suinn (1975), Gomes-Schwartz, Hadley, and Strupp (1978), and Howard and Orlinsky (1972), failed to raise the issue of the clinical relevance of psychotherapy research findings.

Reviewing research on psychotherapeutic processes reported in the late 1960s, Gendlin and Rychlak (1970) echoed two of Ey senck’s earlier themes, lamenting the continuing inadequacy of psychotherapy research methodology—specifically, the paucity of control groups and “properly blind ratings” (p. 156)—and lauding the promise of behavior therapy. Given Cohen’s (1965) then recent comments on significant but meaningless psychotherapy research findings, it is a bit surprising that this review, like later ones during this decade, failed to raise this issue even though most of the research reviewed consists of what would today be called efficacy studies. Moreover, no distinction was drawn in this review between psychotherapy research models analogous to the present-day distinction between efficacy and effectiveness studies; this distinction appeared only later in the decade, presumably because maturation of psychotherapy research methods only then made it meaningful.

In an early review of emerging findings on behavior therapy, Krasner (1971) emphasized repeatedly that behavior therapy, though derived from the experimental psychology laboratory, had nonetheless evolved clinically effective procedures for behavior change. Krasner also informed his readers, most of whom were presumably not behavior therapists, that the behavior therapy researcher uses a number of procedures and designs to go about acquiring information. These are: extrapolation from the experimental lab to clinical application; the use of the clinic as a source of hypotheses for laboratory research; the use of research as treatment; the single case as an experimental model; and the individual as his own control. (p. 515)

Here, Krasner was speaking of the fluidity with which behavior therapy traverses the divide between laboratory and consulting room, making clear that although he recognized a potential distinction between the two, the distinction was moot with regard to the evolving behavior therapies. The same theme is echoed by critics of current efforts to identify empirically supported treatments (see, e.g., Beutler, 1998; Goldfried & Wolfe, 1998), who emphasize the ease with which CBT fits the research model used to identify most of the treatments considered empirically supported.

The 1980s: Continued Advances in Methodology and Efficacy; “Clinical Research Has Little or No Influence on Clinical Practice”

Early in the 1980s, however, researchers began to take up the issue of the clinical utility of psychotherapy research with vigor. Likely reasons include acceptance by many of the meta-analytic concept of effect size as a reflection of the comparative impact of different therapies, as well as other methodological advances in psychotherapy research that heightened its apparent worth. When these advances were coupled with the continuing paradox of the paucity of clinicians consulting the journals when planning interventions, the issue was joined.

One of the earliest of those who commented on this paradox was Barlow (1981), writing in a special section of an issue of the Journal of Consulting and Clinical Psychology entitled “Empirical Practice and Realistic Research: New Opportunities for Clinicians.” Barlow’s comments began provocatively:

At present, clinical research has little or no influence on clinical practice. This state of affairs should be particularly distressing to a discipline whose goal over the last 30 years has been to produce professionals who would integrate the methods of science with clinical practice to produce new knowledge. (p. 147)

Reacting to Barlow’s 1981 call for action, Sargent and Cohen (1983) designed a survey of psychologists to determine the variables associated with the persuasiveness of psychotherapy research strategies for psychologists interested in psychotherapy, strongly suggesting that psychology had finally begun to attend seriously to the issue. Responding to some of the same issues, Shapiro and Shapiro (1982, 1983) undertook a meta-analysis of 143 psychotherapy outcome studies published during the previous 5 years in which two or more treatments were compared with a control group. Although the statistical conclusions and internal validity of these studies were found to be “generally satisfactory,” their construct and external validity were “severely limited by . . . unrepresentativeness of clinical practice” (Shapiro & Shapiro, 1983, p. 42).

The same year, Rosenthal (1983) reiterated the distinction between the statistical and social importance of the effects of psychotherapy and commented explicitly on discussions of effect
sizes derived from recently published meta-analyses by Glass, Smith, and their colleagues (Glass, 1980; Smith, 1980; Smith, Glass, & Miller, 1980). He reintroduced an earlier metric by Rosenthal and Rubin (1979, 1982) for "an intuitively appealing general-purpose effect-size display whose interpretation is perfectly transparent, the binomial effect-size display (BESD)" (Rosenthal, 1983, p. 11). The BESD was designed to heighten the real-world clinical utility of psychotherapy research findings. Rosenthal described it as follows:

The following question is addressed by the BESD: What is the effect on the success rate (e.g., survival rate, cure rate, improvement rate, selection rate, etc.) of the institution of a new treatment procedure? It therefore displays the change in success rate (e.g., survival rate, cure rate, improvement rate, selection rate, etc.) attributable to the new treatment procedure. (p. 11)

Rosenthal (1983) went on to demonstrate that substituting the BESD for the traditional variance estimate makes much more clear than the latter did the extent to which a significant change in outcomes has actually taken place. His compelling example showed that a modest change accounted for by only 10% of the variance can actually correspond to a marked increase in success rate from 34% to 66% or a dramatic decrease in death rate from 66% to 34%.

Kazdin (1986), Koss, Butcher, and Strupp (1986), and Parloff (1984) also commented on an emerging consensus: The psychotherapy outcome studies of the 1980s and before did not adequately reflect—or test—psychotherapy as it is described in textbooks or practiced in the real world. Nonetheless, at exceptional variance to this consensus, in a review of psychotherapy research published in the late 1980s, Goldfried, Greenberg, and Marmar (1990) concluded that "the development of methods for demonstrating clinical (in addition to statistical) significance has been one of the major advances in outcome research" (p. 661).

In a summary of efforts during the 1980s to extend empirical findings on psychotherapy to the consulting room, Hans Strupp (1989) asked rhetorically, "Can the practitioner learn from the researcher?" Strupp described the rub of the problem as follows: "Although research has flourished over the past several decades, there appears to be little evidence that clinicians have adopted a more positive view concerning the practical value of psychotherapy research" (p. 717). He went on to remind his readers that, though psychotherapy has generally been found to be helpful, specific techniques have less often been identified as uniquely effective in treating specific disorders. As a consequence, many clinicians have concluded that what they do and how they do it may be less important than the therapeutic relationship they share with the patient. Strupp's suggested solution: "Instead of focusing on disembodied techniques, we must study and seek a better understanding of the human relationship between a particular patient and a particular therapist and of the transactions occurring between them" (p. 717).

Overview of 40 Years of Psychotherapy Research

This section of the review has traced the emerging recognition, over 40 years, of the infrequency with which psychotherapists relied on the work of psychotherapy researchers. During the decades of the 1950s and 1960s, the problem, largely unrecognized, appeared mainly to reflect the inadequacies of clinical research methodologies. Ironically, though, as these methods improved over the decades of the 1970s and 1980s, their clinical utility became even more of an issue (Nathan, 1998). During these 2 decades, attributions of responsibility shifted to the lack of correspondence between the therapies and patients chosen for empirical research and those typically involved in real-world clinical practice. These concerns, of course, are products of an essential feature of the efficacy model, which would be clearly labeled and thoroughly discussed only later.

The 1990s: Movement Toward and Elaboration of the Distinction Between Efficacy and Effectiveness

Concerns about the usefulness of psychotherapy research for clinicians coalesced in the 1990s as a focus on the distinction between psychotherapy outcome studies that focus on efficacy and those that focus on effectiveness. As detailed earlier, efficacy studies maximize internal validity; they are typically carefully controlled psychotherapy outcome inquiries that use diagnostically homogeneous, randomly assigned patient groups, and, often, manualized, time-limited treatments. By contrast, effectiveness studies aim to maximize external validity; generally taking place in real-world clinical settings, they often lack controls, diagnostic homogeneity, and random assignment of patients, in the effort faithfully to reflect real-world clinical realities. This section briefly reviews a few articles that anticipate the efficacy/effectiveness distinction, then considers in greater detail the outpouring of more recent discussions of the distinction.

The Case Formulation Approach (Persons, 1991)

Persons (1991) reviewed much of the recent literature in this area in 1991, emphasizing the field's growing concerns about the external validity of psychotherapy research and concluding that designs of earlier psychotherapy outcome studies were conceptually incompatible with the models of psychotherapy they were designed to evaluate. Persons blamed this situation on assignment of patients to standardized treatments on the basis of diagnosis rather than on assignment of patients to individualized treatments on the basis of theory-driven psychological assessment of each individual. In so doing, she recalled earlier criticisms of nomothetic approaches and advocacy for idiographic ones, including those of Luborsky (1984), Horowitz and his colleagues (1984), Strupp and Binder (1984), and Weiss, Sampson, and the Mount Zion Psychotherapy Research Group (1986).

Persons's (1991) solution to this longstanding problem: "the case formulation approach to psychotherapy research" (p. 102), which requires development of an "assessment-plus-treatment protocol" (p. 102) based directly on the psychotherapeutic model itself. Outcomes of the case formulation approach are to be cast ideographically because assessments are individualized and outcomes of treatment are to be assessed according to each patient's unique set of problems. Although the existence of treatment manuals that scrupulously reflect their therapeutic models and ideographically assess patients' outcomes would resolve the conceptual incompatibility problems of which Persons despairs, she acknowledged that they have their own problems. Manuals interfere with ongoing therapy and therapist spontaneity and make
comparisons across outcomes difficult when they are based on different sets of evaluative criteria. Nonetheless, in this thoughtful review, Persons cogently identified the problem at the outset of the 1990s and considered the pros and cons of an intriguing solution.

Clarifying Clinically Significant Change (Jacobson & Truax, 1991)

In a 1991 article, Jacobson and Truax (1991) proposed to solve another common outcome research problem affecting clinical validity: statistically significant pre- and posttherapy behavior change scores with little apparent meaning clinically. This problem had been identified more than 25 years earlier by Cohen (1965). Although acknowledging that the effect size statistic derived from meta-analysis is an improvement over standard inferential statistics because it reflects the size of the treatment effect, Jacobson and Truax nonetheless lamented the frequent independence of the size of the effect from its clinical significance. To confront the problem, these authors harkened back to their earlier definition of clinically significant change: movement of someone outside the range of the dysfunctional population or into the range of the functional population (Jacobson, Follette, & Revenstorf, 1984). They then proposed three ways to operationalize this definition:

(a) The level of functioning subsequent to therapy should fall outside the range of the dysfunctional population where range is defined as extending to two standard deviations beyond (in the direction of functionality) the mean for that population.

(b) The level of functioning subsequent to therapy should fall within the range of the functional or normal population, where range is defined as within two standard deviations of the mean of that population.

(c) The level of functioning subsequent to therapy places that client closer to the mean of the functional population than it does to the mean of the dysfunctional population. (Jacobson & Truax, 1991, p. 13)

Reanalysis of NIMH-TDCRP Findings (Ogles, Lambert, & Sawyer, 1995)

Because of its sample size, diagnostic homogeneity, and rigorous treatment standards, the NIMH-TDCRP (Elkin, 1994; Elkin et al., 1985, 1989) represented an extremely important comparative study of mental health treatments. The first of its two primary goals was to test the feasibility and usefulness of the collaborative clinical trial model for psychotherapy research. The second was to test the effectiveness of two brief psychotherapies for the treatment of outpatient depression. Outcomes of treatment for 239 clinically depressed patients randomly assigned to one of four treatments at three sites were compared. The treatments included CBT, IPT, imipramine plus clinical management (CM), and placebo plus CM (PLA-CM). Although all four groups, including the PLA-CM group, reported significant improvements in symptoms, the magnitude of differences among the four immediately and 18 months after treatment was small. Accordingly, the clinical significance of the few statistically significant differences among groups was questioned.

In 1995, Ogles, Lambert and Sawyer (1995) reanalyzed these outcome data, using Jacobson and Truax's (1991) criteria for assessing clinical significance. The initial analysis (Elkin et al., 1989) had failed to consider either the reliability of pre- to post-treatment symptom changes or the clinical significance of multiple measures simultaneously. Accordingly, Ogles and his colleagues hypothesized that reanalysis according to the Jacobson and Truax criteria might be more revealing of clinically significant differences across conditions. In fact, the reanalysis yielded the following findings confirming clinical significance:

- a relatively large number of clients receiving treatment for depression make reliable improvements from pretreatment to post-treatment. These changes are not limited to client's self-report of current symptoms of depression but are also evident to clinical judges and on a self-report measure of diverse physical and psychological symptoms. (Ogles, Lambert, & Sawyer, 1995, p. 324)

Finding that three diverse outcome measures in this instance were consonant convinced Ogles and his colleagues of the validity of this method of assessing clinical significance, even though significant differences among the treatment groups were not found overall.

The Consumer Reports Survey: Effectiveness Study or Consumer Satisfaction Survey?

In 1994, 180,000 subscribers to Consumer Reports were asked to respond to a series of questions about their experiences with mental health professionals, physicians, medications, and self-help groups in the largest survey to date of consumers' views on psychotherapy and other mental health treatments ("Mental Health," 1995). Of the more than 7,000 readers who responded with detailed information about the persons from whom they sought help, about 3,000 reported having talked only to friends, family, or clergy, whereas about 4,100 said they had sought out mental health professionals, family doctors, and/or self-help groups. Thirty-seven percent of those who consulted a mental health professional saw a psychologist, 22% a psychiatrist, 14% a social worker, and 9% a marriage counselor. The 4,100 questionnaire respondents were in clear emotional need: Forty-three percent admitted they were in a "very poor" or "fairly poor" emotional state when they sought help.

The survey's principal findings included the following:

1. Level of satisfaction with therapy was equivalent whether respondents saw a social worker, psychologist, or psychiatrist; those who saw a marriage counselor were somewhat less likely to report having benefited from therapy.

2. Respondents who sought therapy from a family doctor reported doing well, but those who saw a mental health professional for more than 6 months reported doing much better.

3. Psychotherapy alone worked as well as combined psychotherapy and pharmacotherapy; while most persons who took prescribed medication found it helpful, many reported side effects.

4. The longer psychotherapy lasted, the more it helped.

5. Respondents who had tried self-help groups, especially Alcoholics Anonymous, felt especially good about the experience.

These findings painted an extremely encouraging picture of the impact of psychotherapy on this group of respondents. However, the study had important limitations. A mail survey of matters of this sensitivity and complexity, a substantially undefined and essentially undiagnosed group of respondents, and outcome questions that focused on generalized improvement rather than specific
outcomes targeted to symptoms all raised concerns about the meaning of these findings. Troubling as well was the absence of an untreated control or comparison group. In the absence of such a group, it is impossible to say whether respondents who reported having benefited from treatment actually had done so or whether instead they had simply experienced a spontaneous remission of their symptoms with the passage of time.

Another shortcoming of the study was its minimal response rate: Only 4,100 respondents of the 180,000 subscribers to the magazine who had been sent the survey reported seeking professional help or joining groups; only 2,900 reported actually consulting a mental health professional. These numbers are low enough to raise the possibility that substantially more subscribers who had benefited from psychotherapy chose to respond to the questionnaire than those who did not, thereby skewing the findings in a positive direction. In the absence of data on the universe of _Consumer Reports_ subscribers who sought professional help for their emotional problems, this serious design problem cannot be dismissed.

Martin Seligman (1995), a consultant to the project, acknowledged the gap between its methodology and that of other contemporary psychotherapy outcome studies. Nonetheless, he concluded that the _Consumer Reports_ survey “complements the (more traditional) efficacy method, and that the best features of these two methods can be combined into a more ideal method that will best provide empirical validation of psychotherapy” (p. 965). Seligman urged his readers to appreciate the difference between efficacy studies (the traditional gold standard for judging psychotherapy outcomes) and effectiveness studies (feasibility and clinical utility in the real world, as epitomized by the _Consumer Reports_ survey).

We believe, to the contrary, that the _Consumer Reports_ study constituted a survey of consumer satisfaction rather than a psychotherapy effectiveness study. In our view, too many essential elements of both efficacy and effectiveness research—including information on the nature and severity of patients’ diagnoses; therapists’ training and experience; form, length, and nature of outcomes of the treatment; and a reliable metric for reflecting therapeutic change—were lacking to justify considering the study one of effectiveness research. The absence of an untreated comparison group and concerns over possible sampling bias add to our conviction that this study ought not to be held up as an exemplar of effectiveness research.

In 1996, the _American Psychologist_ published a series of commentaries on Seligman’s 1995 article and the _Consumer Reports_ survey. Three articles focused on distinctions between efficacy and effectiveness research.

Hollon (1996) compared the standards to which research on psychopharmacologic agents and on psychotherapy is held, as well as the methods by which their effectiveness is typically assessed. He concluded that, although efficacy studies “leave much to be desired,” effectiveness designs are not a panacea, in large part because they cannot substitute “for the randomized controlled clinical trial when it comes to drawing causal inferences about whether psychotherapy (or any other treatment) actually works” (pp. 1029–1030).

Jacobson and Christensen (1996) questioned Seligman’s assessment of the value of the _Consumer Reports_ study on both conceptual and methodological grounds, concluding that the study is so seriously flawed that few conclusions can be drawn from it. Like Howard et al. (1996) suggested moving away from treatment-focused research, concerned as it is with establishing the comparative efficacy and effectiveness of clinical interventions aggregated over groups of patients, altogether. In its place, they favored patient-focused research, which attempts to monitor an individual’s progress over the course of treatment and to provide feedback of this information to the practitioner, supervisor, or case manager. Patient-focused research focuses on both the actual and estimated progress of the patient, on the basis of his or her clinical characteristics, to provide information on the expected effectiveness of treatment as well as on the characteristics of patients whose response to treatment deviates from expectation.

In the final analysis, the _Consumer Reports_ study has to be assessed as an ambitious failure. At the same time, the response it generated, both pro and con, did illuminate some meaningful new distinctions between efficacy and effectiveness research, largely by forcing both supporters and critics of the study to justify their positions more fully.

**Development of Practice Guidelines (1993–Present)**

The development of practice guidelines by the American Psychiatric Association (1993, 1994b, 1995, 1996, 1997) and the Division of Clinical Psychology of the American Psychological Association (Chambless et al., 1996, 1998; Task Force on Promotion and Dissemination of Psychological Procedures, 1995) has predictably generated strong reactions from those who question the premises of the psychotherapy outcome research on which the guidelines rest. Strong exception, for example, was taken by Sol Garfield (1996), a well-known psychotherapy researcher, to the list of empirically validated treatments first published in 1995 by the Division 12 Task Force. A number of his concerns reflect the efficacy–effectiveness distinction, including the distortion to the psychotherapy process he believed the manuals typically used in efficacy studies cause, as well as the incomparability of psychotherapy patients in efficacy studies and those in real-world psychotherapy settings. Others have expressed similar concerns (see, e.g., Fensterheim & Raw, 1996; Goldfried & Wolfe, 1998).

**Template for Developing Guidelines: Interventions for Mental Disorders and Psychosocial Aspects of Physical Disorders (1995)**

In 1995, the same year the Division 12 Task Force published its initial list of empirically validated treatments, the Template for Developing Guidelines: Interventions for Mental Disorders and Psychosocial Aspects of Physical Disorders (American Psychological Association Task Force, 1995) was proposed by an American Psychological Association task force and subsequently adopted as official association policy by the American Psychological Association Council. Developed “to assure comprehensive-
ness and consistency" (p. 7) of practice guidelines, the Template has two features of relevance to our discussion.

The Template sought to ensure that the efficacy of a treatment included in a practice guideline is established not only by randomized clinical trials (RCTs), the gold standard for this purpose, but also in terms of the treatment comparisons made within those trials. Adding the outcome comparison—to no treatment, to nonspecific treatment variables, or to treatments known to be effective—provides an additional powerful evaluative dimension. RCTs that compare an experimental treatment with established, effective treatments clearly offer a more powerful test than those that compare it with nonrobust comparison treatments or with no treatment at all.

The Template also explicitly distinguished between therapeutic efficacy and effectiveness; it used the term clinical utility to describe the latter construct:

- clinical practice guidelines for behavioral health care (should) be constructed on the basis of two simultaneous considerations or “axes.”
- The first is that guidelines take into consideration a rigorous assessment of scientific evidence with the goal of measuring the efficacy of any given intervention (efficacy). The second axis specifies that guidelines consider the applicability and feasibility of the intervention in the local setting where it is to be proffered (clinical utility).

(American Psychological Association Task Force, 1995, p. 1)

Although a series of well-designed studies might establish the efficacy of an intervention, unless it is effective in real-life clinical settings, it will not be useful:

- The clinical utility axis refers to the ability (and willingness) of practitioners to use, and of patients to accept, the treatment in question, and to the range of applicability of that treatment. It reflects the extent to which the intervention, regardless of the efficacy that may or may not have been demonstrated in the clinical research setting, will be effective in the practice setting in which it is to be applied.

(American Psychological Association Task Force, 1995, p. 13)

Despite its common-sense substance and balanced tone, the Template has also been strongly criticized for many of the same reasons cited by opponents of empirically supported treatments and practice guidelines (as enumerated by Barlow, 1996, and Stricker, 1997).

Another Confirmation of the Dodo Bird Effect (Wampold et al., 1997)

In 1997, Bruce Wampold and his colleagues (Wampold et al., 1997) reported the results of a meta-analysis reexamining the dodo bird effect first proposed by Rosenzweig (1936). Previously reaffirmed by Luborsky, Singer, and Luborsky (1975), Smith and Glass (1977), Stiles, Shapiro, and Elliott (1986), and a number of others but disputed by Krasner (1971), Bergin and Suinn (1975), Rachman and Wilson (1980), and many others (most of them cognitive–behavioral therapists), the dodo bird effect is named for the well-known race in Alice in Wonderland (Carroll, 1865/1962). The dodo bird effect refers to efficacy comparisons among psychotherapies, usually by meta-analysis, that find no differences among them. The meta-analysis by Wampold and his colleagues revealed no differences in efficacy when 277 comparisons among diverse treatments were made, leading these researchers to conclude that there are no meaningful differences in effectiveness among credible (“bona fide”) therapies.

The relevance of the dodo bird effect to the aims of this review is substantial. If there really are no differences in efficacy among psychotherapies, then little purpose is served in seeking to determine whether efficacy or effectiveness studies of psychotherapy provide the clearer picture of the ultimate value of a treatment.

Cris-Christoph (1997) and Howard, Krause, Saunders, and Kopta (1997) criticized the Wampold et al. (1997) meta-analysis on grounds similar to those chosen by earlier critics of earlier meta-analyses of psychotherapy outcomes affirming the dodo bird effect. Above all, they questioned the appropriateness of meta-analysis as a means of exploring differences in psychotherapy outcome and expressed serious concerns about the heterogeneity in design and methodology of the studies chosen for inclusion in the meta-analysis.

So far as we are concerned, the question of whether there are differences in effectiveness among psychotherapies remains open. Despite many efforts through the years to confirm or deny the dodo bird effect once and for all, it has neither been confirmed nor denied. Substantial questions have been raised throughout the years about meta-analyses of psychotherapy outcomes that found no differences in efficacy and, thus, affirmed the dodo bird effect. A growing number of sophisticated RCTs of psychotherapies have in recent years reported differences in psychotherapy effectiveness and, thus, denied the dodo bird effect. Hence, we conclude there is ample justification for trying to determine whether efficacy or effectiveness studies give the truest picture of psychotherapy outcomes, both for its own sake and as a possible alternative means of examining the dodo bird effect from a different perspective.

Review of Contemporary Research on Efficacy and Effectiveness (Kopta, Luenger, Saunders, & Howard, 1999)

Commencing a four-page discussion of efficacy and effectiveness research in a lengthy review of contemporary psychotherapy research, Kopta, Luenger, Saunders, and Howard (1999) acknowledged the continuing gap between clinical research and clinical practice. They suggested that part of the problem might lie in RCTs, proposing instead “that this approach should be replaced by naturalistic designs, which can provide results more applicable to real clinical practice, therefore strengthening external validity” (p. 449).

Like Seligman (1996), Howard et al. (1996), and Wampold (1997), Kopta et al. (1999) endorsed effectiveness studies as the best means of understanding psychotherapy's impact in real-world clinical settings. This position derives in part from research on Howard's dosage model (Howard, Kopta, Krause, & Orlinsky, 1986). However, because dose–effect designs, including RCTs, use grouped data and provide information about average patients, they are not as useful to the psychotherapist as they could be. Accordingly, Howard, Orlinsky, and Luenger (1995) and Tingey, Lambert, Burlinahme, and Hansen (1996) proposed a dose–outcome design, a single-case application of the dosage model, which tracks therapeutic progress across sessions. Howard et al. (1996) also proposed patient profiling, which would track improvement from pretreatment clinical status by means of hierarchical linear modeling. The aim of both is to increase the utility of
data from effectiveness studies by enhancing the reliability of their measurement systems.

Kopka et al. (1999) also noted the increasing emphasis psychotherapy researchers have recently put on the development of new approaches to assessment. To this end, they described creative RCTs that have been used or proposed that can distinguish active ingredients if indeed they exist, dose-effect studies (that) can discover lawful outcome relationships across sessions for patients treated by different therapies as practiced in clinical settings, using dose-outcome designs to group individual patients by treatment type and similar dose-response patterns... to answer Paul’s patient-focused question, (and) outcome studies... to distinguish which psychotherapies are more efficient in addition to which ones are simply effective. (p. 453)

Current Efforts to Integrate Efficacy and Effectiveness Studies

As the foregoing indicates, the evolution of concerns about the usefulness of psychotherapy research for clinical practice has progressed substantially from Eysenck’s time. For the most part, these changes have occurred in step with advances in psychotherapy research methodology and the robustness of psychotherapy interventions.

During the 1950s and 1960s, although some psychologists spoke and wrote of their concerns that statistical significance is not always associated with clinical worth, little was said about the impact of psychotherapy research on the practice of psychotherapy, perhaps because of the primitive state of the methodology of psychotherapy research. Though the 1970s witnessed marked advances in research methodology that paralleled the emergence of a variety of promising behavioral and cognitive-behavioral techniques, surprisingly little was written about how these advances might affect clinical practice.

Early in the 1980s, however, researchers began to dwell in earnest on the usefulness of psychotherapy research to clinicians. The developments in research methodology of the 1970s played a role in the emergence of these concerns, both because they enabled researchers to evaluate the comparative impact of different therapies and because the worth of psychotherapy research increasingly began to be appreciated. The solutions proposed to resolve the problem during this time generally focused on efforts to make efficacy studies more relevant to clinical practice, though a few commentators systematically weighed the respective contributions of what today is called efficacy and effectiveness research. Some researchers also focused the issue by suggesting that research on psychotherapy effectiveness might have a less substantial ultimate payoff than a focus on common variables.

These more general concerns about the usefulness of psychotherapy research coalesced early in the decade of the 1990s around the meaning and significance of distinctions between psychotherapy outcome studies that focus on efficacy and those that focus on effectiveness. Some researchers have claimed that improvements in efficacy research would be sufficient to fix the problem, others have emphasized effectiveness studies as an exclusive answer, and still others have recommended simultaneous attention to both. This brings us to the most recent period in this evolution, the latter half of the decade of the 1990s, a period characterized by initial efforts to integrate efficacy and effectiveness research into a coherent whole. A recent issue of the American Psychological Association’s new electronic journal Prevention & Treatment featured three articles proposing two different approaches to this end.

The NIMH Integration Initiative

Norquist, Lebowitz, and Hyman (1999)—(the last being the current director of NIMH) began their article by acknowledging that “the intrinsic efficacy of an intervention (either pharmacological or psychotherapeutic) is not usually informative for treatment practice in the community” (abstract). The problem lies in the difference in the yield of research that follows a regulatory model (based on the requirements of the federal Food and Drug Administration) and research that adheres to a public health model (in which research is designed to evaluate the effectiveness of clinical interventions as they are likely to be delivered in community and specialized practice). The article’s bottom line is a proposal for how the NIMH, in consultation with basic scientists, advocates, and other federal agencies, might bridge the gap between the regulatory (efficacy) and the public health (effectiveness) models.

Norquist et al. (1999) proposed a new paradigm. It would incorporate both experimental and observational work, albeit after changes in methods for both. Although their ideas are clearly in the process of evolution, Norquist and his colleagues suggested the following:

Research designs must permit a loosening of exclusion criteria to allow for enrollment of people with different levels of disease severity and comorbid conditions. In addition, treatment settings must be more diversified to allow for a range of providers from primary care, managed care settings to tertiary care, academic centers. Outcome measures will need to incorporate domains that are important to consumers, families and policy makers (e.g., performance, disability, cost, resource use, etc.)... In addition, specific areas of treatment intervention research need to be launched (e.g., rehabilitation research) and reinvigorated (e.g., psychosocial intervention research). (¶ 16–17)

Research according to this new paradigm must “combine the designs of traditional clinical and services research studies” (Norquist et al., 1999, ¶ 18). Doing so requires compromises between the strict randomized designs of traditional clinical research and the more flexible observational designs of services research. Merging these designs requires NIMH

... to bring together methodologists with expertise across these fields to delineate what we currently know and what we don’t (because it is) quite likely that new methods and statistical analytic approaches will need to be developed to address studies in the mental health area. (¶ 18)

To achieve these ambitious goals, a new paradigm, new methods, and new statistical approaches are all required. New methods of grant review and a new research infrastructure to facilitate submission, review, and funding of research within this new paradigm are also envisioned. It is clear that the entire enterprise must founder unless the new methods and new statistical procedures, which remain to be developed, can in fact be produced.

A companion article, written by Niedererhe, Street, and Lebowitz (1999), detailed the new organizational entities that would facilitate NIMH’s new research emphasis on the public health model
and briefly described design changes researchers would be expected to incorporate in grant proposals that address the new research paradigm. Unfortunately, as the authors themselves observed,

at the present time, it is easier to recognize when a proposed design presents a good fit for the particular questions addressed in a given study than to specify in advance or in the abstract the desired features for any and all public health model studies. (¶ 10)

In other words, specific advice for researchers wishing to focus their research on this new paradigm is not yet available. Nonetheless, plans for several large-scale, multisite clinical trials that incorporate numerous features typically identified with effectiveness research are described. Both psychosocial and psychopharmacological treatments for specific target disorders will be compared in these trials, the first two of which will investigate treatments for bipolar disorder in adults and depression in children.

As noted in the article by Norquist and his coworkers (1999), the Clinical Treatment and Services Research Workgroup of the National Advisory Mental Health Council recently published Bridging Science and Service (National Advisory Mental Health Council Workgroup, 1999). This document contains a “number of recommendations about methods development that NIMH should foster in this area of research, including some that relate to innovative combinations of research designs” (p. 4). Predictably, the report recommended that NIMH continue its support of efficacy, effectiveness, practice, and service systems research and expend additional effort to build bridges across these research foci. Examples of bridge-building across these domains include loosening restrictive exclusion criteria and increasing the range of outcome measures in efficacy studies, seeking cost-effectiveness data and studying commonly practiced but not commonly studied interventions in effectiveness studies, and putting more resources into clinical epidemiology and dissemination research.

The report (National Advisory Mental Health Council Workshop, 1999) also recommended that “NIMH should explore new methods for analysis of data from studies that incorporate innovative combinations of research designs” (p. 46) and “NIMH should encourage development of methods to explicitly evaluate trade-offs in alternate design features that differ in their implications for internal and external validity” (p. 47). In the first instance, the report envisioned studies that combine both experimental and observational designs or analytic features to address questions about treatment in a community context. These studies would be of longer duration, larger sample sizes, and broader domains of outcome than at present. As a result, a reduction in the intensity (or depth) of data collection may be required for such a study to be feasible. Such alternative approaches might entail a trade-off between breadth and depth of information collected and analyzed on the one hand, and the ability of studies to inform either clinical or policy audiences on the other. “Methodologists should develop approaches to model those trade-offs in breadth and depth for given study purposes and to maintain reasonable costs for the study” (p. 46).

In the second instance, innovative methods, otherwise unspecified, need to be fostered:

Pursuit of the goals of methodological development requires integration of clinical and social science, particularly incorporation of statistical and econometric expertise. Expertise in qualitative analyses also is likely to be needed to generate new approaches to studying systems, treatments, and adherence with treatment. (National Advisory Mental Health Council Workgroup, 1999, p. 47)

In summary, the NIMH proposal embraced a dual strategy of modifying efficacy and effectiveness research on specific issues to maximize real-world pay-off and developing new research designs and statistical procedures to avoid or minimize the well-known problems associated with efforts to adjust internal and external validity simultaneously. “When studies involve multiple design decisions that favor external or internal validity, rarely are both improved simultaneously” (National Advisory Mental Health Council Workgroup, 1999, p. 46). Whether these ambitious goals will be realized, of course, is anyone’s guess at the present time.

**Multisite Efficacy/Effectiveness Clinics**

Klein and Smith (1999) proposed the development of “dedicated, multisite efficacy/effectiveness clinics” (abstract) to address the problems posed by the conflicting demands of internal and external validity in efficacy and effectiveness studies. These special clinics would also be structured to examine such understudied treatment issues as compliance, comorbidity, refractory illness, and withdrawal syndromes, as well as adjunctive and maintenance treatments. The clinics would emphasize careful, reliable, documented studies on process and outcome. They would also promote development of outcome norms for well-defined populations on such variables as diagnosis, economic status, history, and comorbidity. The proximate goal would be to generate, across cooperating clinics, “a large volume of well-delineated patients [who] could be treated and studied who may have high comorbidity with medical, psychiatric, and substance abuse conditions” (¶ 19). The distal goals would be both to develop benchmarks for expected treatment outcomes for these distinct groups of patients by means of normative sampling and to serve as hypothesis-generating therapeutic endeavors.

As with NIMH’s new paradigm, however, Klein and Smith’s (1999) proposal for efficacy/effectiveness clinics is long on enthusiasm, problem identification, and aspirations for change, but a good deal shorter on concrete details of design, methodology, and statistical analysis. Although understandable at this stage in the development of the concept, the relative lack of substance leaves the reader who appreciates some of the problems attendant on integrating efficacy and effectiveness studies uncertain to what extent integration can actually be effected. Instead, one is left to wonder whether these clinics would function largely as a more efficient means of meeting an older goal, strengthening the sequence linking initial hypothesis testing in efficacy studies and confirmation of these hypotheses by effectiveness studies.

**Meta-Analyses of Psychotherapy in Clinically Representative Conditions**

The NIMH effort to integrate efficacy and effectiveness research (National Advisory Mental Health Council Workgroup, 1999; Niedereehe, Street, & Lebowitz, 1999; Norquist, Lebowitz, & Hyman, 1999) assumed that these two research models, separately and jointly, ultimately lack the capacity to optimize internal and external validity. The Klein and Smith (1999) model of multisite
efficacy/effectiveness clinics appears to draw the same conclusion, even though the proposal for solving the problem by this means differs somewhat from the NIMH proposal.

Recent analyses of the psychotherapy outcome literature by Shadish and his colleagues (Shadish, Matt, Navarro, & Phillips, in press; Shadish et al., 1993, 1997) have taken a very different starting point but addressed a closely related set of issues. Using sophisticated random effects regression analyses and secondary analyses of prior meta-analyses, Shadish and his associates asked whether psychotherapy outcome studies ranging from less to more clinically representative differ in effectiveness, as reflected by a large array of outcome variables. In other words, they asked whether there are outcome differences in psychotherapy completed for efficacy research purposes and psychotherapy completed for effectiveness research purposes, thereby searching for substantive differences in results as a function of substantive differences in research methods.

Shadish and his colleagues consistently failed to find differences in efficacy/effectiveness as a function of where on the efficacy/effectiveness (clinically representative) continuum a study falls. Although this finding does not reduce the real methodological differences between the efficacy/effectiveness research models to semantic ones, it does suggest that, in terms of one very important factor, clinical usefulness, the distinction may be more apparent than real. This research also confirms prior findings indicating that therapy is more effective in larger doses when outcome measures are highly tailored to treatment and, less consistently, when behaviorally oriented therapies are used. These latter results buttress the view that the innovative data-analytic procedures used by this research team are sensitive to crucial variables affecting outcome.

In the first of these articles, Shadish and his colleagues (1997) asked 15 meta-analysts to provide effect sizes from 56 studies included in previous reviews that met one of three increasingly stringent levels of clinical representativeness. Effect sizes were then synthesized and compared with results from the original meta-analyses. Results indicated that the effect sizes of the more and the less clinically representative studies were essentially the same at all three criteria levels. Shadish et al. (in press) subsequently confronted some of the methodological problems of the 1997 study by synthesizing results from 90 psychotherapy outcomes that ranged widely in clinical representativeness. Random effects regression analyses indicated that the psychotherapies were equally effective across the clinical representativeness continuum, thereby confirming the 1997 findings.

Shadish's research calls into question the implicit assumption that efficacy and effectiveness studies of the same treatment modality yield incomparable outcomes as a function of their design differences. They suggest that these differences in design may be more apparent than real, at least in terms of their capacity to test the robustness of a treatment.

In Conclusion

The discordant literature on efficacy and effectiveness research reviewed here parallels the continuing lack of consensus among psychotherapy researchers themselves and between them and clinicians on how best to add value to research findings on psychotherapy for practitioners. Some investigators continue to believe that the designs of clinical trials and other efficacy studies require neither substantial change nor enhancement with data from effectiveness studies. Others as strongly believe that there is a problem with exclusive reliance on the efficacy model. Of this group, some would rely almost exclusively on effectiveness studies, whereas others would work to bring efficacy and effectiveness research models into greater harmony with one another. Still others conclude that, regardless of novel adjustments and new configurations, effectiveness and efficacy studies will never provide clinicians the information they require to do their jobs better; these individuals opt for continuing to seek a new paradigm altogether. Despite almost a decade of research on the efficacy/effectiveness distinction, a metric permitting rational choice between these alternatives has not yet been developed. Adding to the problem is the absence of agreement on the essential components of either efficacy or effectiveness trials: Seligman (1995, 1996) has been severely criticized for proclaiming the Consumer Reports survey a model effectiveness study rather than, as many have seen it, an equivocal survey of consumer satisfaction.

As troublesome as the continuing lack of consensus on an optimal therapy outcome research model after more than 50 years of efforts to develop such a model is the continuing gap in understanding between psychotherapy researchers and psychotherapists. A rift that has existed since Eysenck's (1952) initial review of psychotherapy outcomes so incensed the clinicians of his time, the inability of researchers to generate findings of value to clinicians continues to have a serious impact on the credibility of research findings on outcomes (Nathan, 1999). A substantial part of the problem, of course, is the inability of researchers to agree on either a most appropriate research model or the meaning of data from existing models. An apt illustration is the continuing inability of researchers to resolve the dodo bird conjecture, epitomized by the 1997 meta-analysis of psychotherapy outcomes by Wampold and his colleagues (1997) and its accompanying critiques. If behavioral scientists cannot reliably document differences in effectiveness among treatments that clinicians are convinced do exist, why should clinicians put any faith in the efforts of researchers?

If this half-century dilemma is to be resolved anytime soon, it will require agreement on definitions of and the best means for using both efficacy and effectiveness studies. The APA Template for Developing Guidelines (American Psychological Association Task Force, 1995) might be a model for such an effort, despite the decidedly cool reception it has received from the practice community thus far. Despite their appeal to the quantitative mind, meta-analyses of psychotherapy outcomes seem likely to continue to generate irresolvable disagreement, as they have for more than a quarter of a century. Pending outcomes from the massive NIMH effort to integrate efficacy and effectiveness studies by means of new methodologies and statistics that have yet to be developed, continued efforts to improve the designs of efficacy and effectiveness research must be made. Ground rules for moving from efficacy studies to effectiveness studies need to be developed: When do practitioners know enough about a treatment's worth in the ideal environment of the efficacy study to determine its effectiveness with larger, more diverse populations in real-world settings? We are encouraged to believe that a focus on efficacy and effectiveness research may well resolve this problem of 50 years standing, given how far we have come in the development of these approaches in less than a decade.
References


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


Mintz, J., Drake, R. E., & Crits-Cristoph, P. (1996). Efficacy and effect-


Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically-validated psycho-


Received June 18, 1999
Revision received January 5, 2000
Accepted January 5, 2000